Chapter 12
20 Years of Learning About Vision: Questions Answered, Questions Unanswered, and Questions Not Yet Asked

Bruno A. Olshausen

Abstract I have been asked to review the progress that computational neuroscience has made over the past 20 years in understanding how vision works. In reflecting on this question, I come to the conclusion that perhaps the most important advance we have made is in gaining a deeper appreciation of the magnitude of the problem before us. While there has been steady progress in our understanding—and I will review some highlights here—we are still confronted with profound mysteries about how visual systems work. These are not just mysteries about biology, but also about the *general principles* that enable vision in any system whether it be biological or machine. I devote much of this chapter to examining these open questions, as they are crucial in guiding and motivating current efforts. Finally, I shall argue that the biggest mysteries are likely to be ones we are not currently aware of, and that bearing this in mind is important as it encourages a more exploratory, as opposed to strictly hypothesis-driven, approach.

Introduction

I am both honored and delighted to speak at this symposium. The CNS meetings were pivotal to my own coming of age as a scientist in the early 1990s, and today they continue to constitute an important part of my scientific community. Now that 20 years have passed since the first meeting, we are here today to ask, what have we

B.A. Olshausen (✉)
Helen Wills Neuroscience Institute and School of Optometry, University of California—Berkeley, Berkeley, CA, USA

Redwood Center for Theoretical Neuroscience, UC Berkeley, 575A Evans Hall, MC 3198, Berkeley, CA 94720-3198, USA
e-mail: baolshausen@berkeley.edu

J.M. Bower (ed.), *20 Years of Computational Neuroscience*, Springer Series in Computational Neuroscience 9, DOI 10.1007/978-1-4614-1424-7_12,

learned? I have been tasked with addressing the topic of vision, which is of course a huge field, and so before answering I should disclose my own biases and the particular lens through which I view things: I began as an engineer wanting to build robotic vision systems inspired by biology, and I evolved into a neuroscientist trying to understand how brains work inspired by principles from mathematics and engineering. Along the way, I was fortunate to have worked and trained with some of the most creative and pioneering scientists of our field: Pentti Kanerva, David Van Essen, Charlie Anderson, Mike Lewicki, David Field, and Charlie Gray. Their own way of thinking about computation and the brain has shaped much of my own outlook, and the opinions expressed below stem in large part from their influence. I also benefited enormously from my fellow students in the Computation and Neural Systems program at Caltech in the early 1990s and the interdisciplinary culture that flourished there. They impressed upon me that the principles of vision are not owned by biology, nor by engineering—they are universals that transcend discipline, and they will be discovered by thinking outside the box.

Now to begin our journey into the past 20 years, let us first gain some perspective by looking back nearly half a century, to a time when it was thought that vision would be a fairly straightforward problem. In 1966, the MIT AI Lab assigned their summer students the task of building an artificial vision system (Papert 1966). This effort came on the heels of some early successes in artificial intelligence in which it was shown that computers could solve simple puzzles and prove elementary theorems. There was a sense of optimism among AI researchers at the time that they were conquering the foundations of intelligence (Dreyfus and Dreyfus 1988). Vision it seemed would be a matter of feeding the output of a camera to the computer, extracting edges, and performing a series of logical operations. They were soon to realize however that the problem is orders of magnitude more difficult. David Marr summarized the situation as follows:

> …in the 1960s almost no one realized that machine vision was difficult. The field had to go through the same experience as the machine translation field did in its fiascoes of the 1950s before it was at last realized that here were some problems that had to be taken seriously. … the idea that extracting edges and lines from images might be at all difficult simply did not occur to those who had not tried to do it. It turned out to be an elusive problem. Edges that are of critical importance from a three-dimensional point of view often cannot be found at all by looking at the intensity changes in an image. Any kind of textured image gives a multitude of noisy edge segments; variations in reflectance and illumination cause no end of trouble; and even if an edge has a clear existence at one point, it is as likely as not to fade out quite soon, appearing only in patches along its length in the image. The common and almost despairing feeling of the early investigators like B.K.P. Horn and T.O. Binford was that practically anything could happen in an image and furthermore that practically everything did. (Marr 1982)

The important lesson from these early efforts is that it was from *trying to solve the problem* that these early researchers learned what were the difficult computational problems of vision, and thus what were the important questions to ask. This is still true today: Reasoning from first principles and introspection, while immensely valuable, can only go so far in forming hypotheses that guide our study of the visual system. *We will learn what questions to ask by trying to solve the problems of vision.* Indeed,

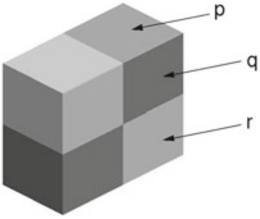

Fig. 12.1 Image of a block painted in two shades of *gray* (from Adelson 2000). The edges in this image are easy to extract, but understanding what they mean is far more difficult

this is one of the most important contributions that computational neuroscience can make to the study of vision.

A decade after the AI Lab effort, David Marr began asking very basic questions about information processing in the visual system that had not yet been asked. He sought to develop a computational theory of biological vision, and he stressed the importance of *representation* and the different types of information that need to be extracted from images. Marr envisioned the problem being broken up into a series of processing stages: a primal sketch in which features and tokens are extracted from the image, a 2.5D sketch that begins to make explicit aspects of depth and surface structure, and finally an object-centered, 3D model representation of objects (Marr 1982). He attempted to specify the types of computations involved in each of these steps as well as their neural implementations.

One issue that appears to have escaped Marr at the time is the importance of *inferential computations* in perception. Marr's framework centered around a mostly feedforward chain of processing in which features are extracted from the image and progressively built up into representations of objects through a logical chain of computations in which information flows from one stage to the next. After decades of research following Marr's early proposals, it is now widely recognized (though still not universally agreed upon) by those in the computational vision community that the features of the *world* (not images) that we care about can almost never be computed in a purely bottom-up manner. Rather, they require inferential computation in which data is combined with prior knowledge in order to estimate the underlying causes of a scene (Mumford 1994; Knill and Richards 1996; Rao et al. 2002; Kersten et al. 2004). This is due to the fact that natural images are full of ambiguity. The causal properties of images—illumination, surface geometry, reflectance (material properties), and so forth—are entangled in complex relationships among pixel values. In order to tease these apart, aspects of scene structure must be estimated simultaneously, and the inference of one variable affects the other. This area of research is still in its infancy and models for solving these types of problems are just beginning to emerge (Tappen et al. 2005; Barron and Malik 2012; Cadieu and Olshausen 2012). As they do, they prompt us to ask new questions about how visual systems work.

To give a concrete example, consider the simple image of a block painted in two shades of gray, as shown in Fig. 12.1 (Adelson 2000). The edges in this image are easy to extract, but understanding what they mean is far more difficult. Note that there are three different types of edges: (1) those due to a change in reflectance (the boundary between *q* and *r*), (2) those due to a change in object shape (the boundary between *p* and *q*), and (3) those due to the boundary between the object and background. Obviously it is impossible for any computation based on purely local image analysis to tell these edges apart. It is the context that informs us what these different edges mean, but how exactly? More importantly, *how are these different edges represented in the visual system and at what stage of processing do they become distinct?*

As one begins asking these questions, an even more troubling question arises: How can we not have the answers after a half century of intensive investigation of the visual system? By now there are literally mounds of papers examining how neurons in the retina, LGN, and V1 respond to test stimuli such as isolated spots, white noise patterns, gratings, and gratings surrounded by other gratings. We know much—perhaps too much—about the orientation tuning of V1 neurons. Yet we remain ignorant of how this very basic and fundamental aspect of scene structure is represented in the system. The reason for our ignorance is not that many have looked and the answer proved to be too elusive. Surprisingly, upon examining the literature one finds that, other than a handful of studies (Rossi et al. 1996; Lee et al. 2002; Boyaci et al. 2007), no one has bothered to ask the question.

Vision, though a seemingly simple act, presents us with profound computational problems. Even stating what these problems are has proven to be a challenge. One might hope that we could gain insight from studying biological vision systems, but this approach is plagued with its own problems: Nervous systems are composed of many tiny, interacting devices that are difficult to penetrate. The closer one looks, the more complexity one is confronted with. The solutions nature has devised will not reveal themselves easily, but as we shall see the situation is not hopeless.

Here I begin by reviewing some of the areas where our field has made remarkable progress over the past 20 years. I then turn to the open problems that lie ahead, where I believe we have the most to learn over the next several decades. Undoubtedly though there are other problems lurking that we are not even aware of, questions that have not yet been asked. I conclude by asking how we can best increase our awareness of these questions, as these will drive the future paths of investigation.

Questions Answered

Since few questions in biology can be answered with certainty, I cannot truly claim that we have fully answered any of the questions below. Nevertheless these are areas where our field has made concrete progress over the past 20 years, both in terms of theory and in terms of empirical findings that have changed the theoretical landscape.

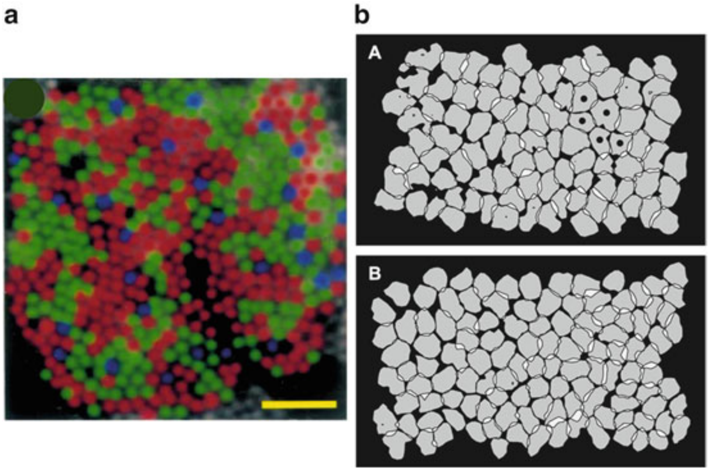

Fig. 12.2 Tiling in the retina. (**a**) Tiling of L, M, S cones; scale bar = 5 arcmin (from Roorda and Williams 1999). (**b**) Tiling of parasol retinal ganglion cell receptive fields; *A*, on cells; *B*, off cells (from Gauthier et al. 2009a, b)

Tiling in the Retina

A long-standing challenge facing computational neuroscience, especially at the systems level, is that the data one is constrained to work with are often sparse or incomplete. Recordings from one or a few units out of a population of thousands of interconnected neurons, while suggestive, cannot help but leave one unsatisfied when attempting to test or form hypotheses about what the system is doing as a whole. In recent years, however, a number of advances have made it possible to break through this barrier in the retina.

The retina contains an array of photoreceptors of different types, and the output of the retina is conveyed by an array of ganglion cells which come in even more varieties. How these different cell types tile the retina—that is, how a complete population of cells of each type cover the two-dimensional image through the spatial arrangement of their receptive fields—has until recently evaded direct observation. As the result of advances in adaptive optics and multielectrode recording arrays, we now have a more complete and detailed picture of tiling in the retina which illuminates our understanding of the first steps in visual processing.

Adaptive optics corrects for optical aberrations of the eye by measuring and compensating for wavefront distortions (Roorda 2011). With this technology, it is now possible to resolve individual cones within the living human eye, producing breathtakingly detailed pictures of how L, M, and S cones tile the retina (Fig. 12.2a) (Roorda and Williams 1999). Surprisingly, L and M cones appear to be spatially clustered beyond what one would expect from a strictly stochastic positioning according to density (Hofer et al. 2005). New insights into the mechanism of color perception have been obtained by stimulating individual cones and looking at how subjects report the corresponding color (Hofer and Williams 2005). Through computational modeling studies, one can show that an individual cone's response is interpreted according to a Bayesian estimator that is attempting to infer the actual color present in the scene in the face of subsampling by the cone mosaic, not simply the cone's "best color" (Brainard et al. 2008). It is also possible to map out receptive fields of LGN neurons cone by cone, providing a more direct picture of how these neurons integrate across space and wavelength (Sincich et al. 2009).

Another important question that can be addressed with adaptive optics is the effect of fixational drifts and microsaccades on perception. It is now possible to track movements of the retina in real-time with single-cone precision, allowing one to completely stabilize retinal images or even introduce artificially generated drifts (Vogel et al. 2006; Arathorn et al. 2007). These studies strongly suggest the presence of internal mechanisms that compensate for drifts during fixation to produce stable percepts (Austin Roorda, personal communication).

At the level of retinal ganglion cells, large-scale neural recording arrays have enabled the simultaneous mapping of receptive fields over an entire local population (Litke et al. 2004). These studies reveal a beautifully ordered arrangement not only in how receptive fields are positioned but also in how they are shaped so as to obtain optimal coverage of the image for each of the four major cell types (i.e., each of the different combinations of on/off and midget/parasol) (Gauthier et al. 2009a, b). Although the position of receptive fields can be somewhat irregular, the shape of each receptive field is morphed so as to fill any gaps in coverage, as shown in Fig. 12.2b. Remarkably, despite the irregular spacing, the receptive field overlap with nearest neighbors is fairly constant, which is a further testament to the degree of precision that is present in retinal image encoding.

Together, these developments provide a solid picture of retinal organization and resolve questions regarding the completeness of coverage that were unresolved just a decade ago. Importantly, these developments also open a new door in allowing us to ask more detailed questions about the link between neural mechanisms and perception.

The Relation Between Natural Image Statistics and Neural Coding

Twenty years ago, most people (myself included) thought of neurons at early stages of the visual system in terms of feature detection. For example, Marr had proposed that retinal ganglion cells function as edge detectors by computing zero crossings of the Laplacian operator (which indicates extrema in the first derivative) and this became a fairly popular idea. Similarly, the oriented receptive fields of V1 neurons were thought to operate as oriented edge detectors that encode the boundaries or geometric shape of objects. However, in the early 1990s it became clear there is another way to think about what these neurons are doing in terms of *efficient coding principles*. Here the goal is to consider how information about the image can be encoded and represented in a complete manner that is adapted to the input statistics. In contrast to detection, which is typically a lossy process designed for a specific purpose, the goal of efficient coding is to form a generic representation that could be used for myriad tasks, but which nevertheless exploits and makes explicit the structure contained in natural images.

Although the efficient coding hypothesis was first proposed by Barlow more than 50 years ago (Barlow 1961), it was not until decades later that investigators such as Laughlin and Srinivasan began making serious quantitative connections between the statistics of natural scenes and neural coding (Srinivasan et al. 1982). David Field subsequently showed that the power spectrum of natural images follows a characteristic $1/f^2$ power law, and he pointed out how the scale-invariant structure of cortical receptive fields is well matched to encode this structure (Field 1987). Atick and Redlich formulated the whitening theory of retinal coding, which proposed that the purpose of the circularly symmetric, center-surround receptive fields of retinal ganglion cells is not to detect edges as Marr claimed, but rather to remove redundancies in natural images so as to make maximal use of channel capacity in the optic nerve (Atick and Redlich 1992). Subsequent neurophysiological experiments in the LGN seemed to support this assertion (Dan et al. 1996). Around the same time, David Field and I showed through computer simulation that the localized, oriented, and multiscale receptive fields of V1 neurons could be accounted for in terms of a sparse coding strategy adapted to natural images (Olshausen and Field 1996). These theories and findings have drawn considerable interest because they offer an intimate, quantitative link between theories of neural coding and experimental data. Moreover it is not just a theory of vision, but a general theory of sensory coding that could be applied to other modalities or subsequent levels of representation, and indeed there has been much work investigating these directions (Geisler et al. 2001; Hyvarinen and Hoyer 2001; Schwartz and Simoncelli 2001; Karklin and Lewicki 2003, 2005, 2009; Hyvarinen et al. 2005; Smith and Lewicki 2006).

A related theoretical framework that has been used to make connections between natural scene statistics and neural representation is that of *Bayesian inference*. Here the goal is to go beyond coding to consider how the properties of scenes are inferred from image data. As mentioned above, making inferences about the world depends upon strong prior knowledge. Often this knowledge is probabilistic in nature. For example, in the simple scene of Fig. 12.1, we could choose to interpret it either as a flat scene created entirely by paint (which it is), as a scene created entirely by structured light, or as a three-dimensional object in two shades of paint (Adelson 2000). All three are valid interpretations when judged purely in terms of the image data. Our visual system chooses the latter interpretation because it is the most parsimonious or *probable* interpretation that is consistent not only with the data but also with our experience in interacting with the world. A goal of many modeling efforts over the past 20 years has been to show how probabilistic information about the world can be learned from visual experience and how inferential computations can be performed in neural systems (Dayan et al. 1995; Rao et al. 2002; Ma et al. 2006). Some of these models make predictions about higher level visual representations beyond V1, in addition to providing a possible account for the role of feedback connections from higher areas to lower areas (Lee and Mumford 2003; Karklin and Lewicki 2005; Cadieu and Olshausen 2012). An important property of these models is the manner in which different hypotheses compete to explain the data—termed "explaining away" (Pearl 1988)—which provides an account for the nonlinear, suppressive effects of context upon the responses of visual neurons (Vinje and Gallant 2000; Murray et al. 2002; Zhu and Rozell 2011).

The Nature of Intermediate-Level Vision

For many years intermediate-level vision was the *terra incognita* of our field. It is the murkiest territory because unlike low-level vision its neural substrates cannot be directly identified or characterized, and unlike high-level phenomena such as object recognition and attention we have no well-established terms or conceptual frameworks for what goes on at this stage. In fact, it is difficult even to define what "intermediate-level vision" means. Processes such as grouping or segmentation are often ascribed to this stage, but the range of other things that could be going on is so broad and ill-defined that it is semi-seriously referred to as "everything between low-level and high-level vision." Over the past 20 years however this area has become progressively less murky through insightful and penetrating psychophysical experiments.

In particular, Nakayama and colleagues have provided compelling evidence that intermediate-level representations are organized around *surfaces* in the 3D environment, and that these representations serve as a basis for high-level processes such as visual search and attention (Nakayama et al. 1995). This view stands in contrast to previous theories postulating 2D features such as orientation and motion energy as the basis of perceptual grouping that underlies texture segmentation, search, and attention (Treisman and Gelade 1980; Julesz 1981). Nakayama's experiments suggest that representations of 3D surface structure are formed prior to this stage, and that perceptual grouping operates primarily on surface representations rather than 2D features. For example, when colored items are arranged on surfaces in different depth planes, detection of an odd-colored target is facilitated when pre-cued to the depth plane containing the target; but if the items are arranged so as to appear attached to a common surface receding in depth, then pre-cueing to a specific depth has little effect. Thus, it would appear that attention spreads within surfaces in 3D coordinates in the environment, not within 2D proximity or a simple disparity measure.

Another contribution of Nakayama's work is in pointing out the importance of *occlusion* in determining how features group within a scene. Once again, they show that simple grouping rules based on 2D proximity or similarity do not suffice. This should not be surprising, because under natural viewing conditions the 2D image

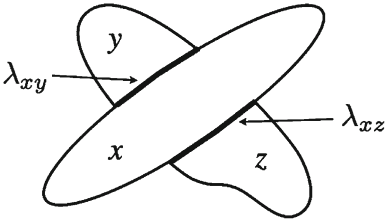

Fig. 12.3 Occlusion and border ownership. When image regions corresponding to different surfaces meet in the projection of a scene, the region corresponding to the surface in front "owns" the border between them. A region that does not own a border is essentially unbounded and can group together with other unbounded regions. Here, surface x owns the borders λ_{xy} and λ_{xz}. Thus, regions y and z are unbounded at these borders and they are free to group with each other, but not with region x because it owns these borders and is therefore bounded by them (adapted from Nakayama et al. 1995)

arises from the projection of 3D surfaces in the environment. When these surfaces overlap in the projection, the one nearest the observer "overwrites" or occludes the other. Thus, a proper grouping of features would need to take this aspect of scene composition into account in determining what goes together with what, as shown in Fig. 12.3. By manipulating disparity cues so as to reverse figure–ground relationships in a scene, they show that the visual system groups features in a way that obeys the rules of 3D scene composition. Features are grouped within surfaces, even when parts of the surface are not visible, but not beyond the boundary of a surface. Thus, the neural machinery mediating this grouping would seem to require an explicit representation of border ownership, such as described by von der Heydt (Zhou et al. 2000; Qiu and von der Heydt 2005), or some other variable that expresses the boundaries and ordinal relationship of surfaces.

Nakayama's work is not the only in this realm, there are many others (Adelson 1993; Mamassian et al. 1998; Knill and Saunders 2003). It is a body of work that suggests what to look for at the neural level. Much as color psychophysics preceded the discovery of its neural mechanisms, these psychophysical experiments suggest the existence of certain neural representations at the intermediate level of vision.

Functional Organization of Human Visual Cortex

In 1991, Felleman and Van Essen published their now famous diagram of connections between visual cortical areas in the macaque monkey (Felleman and Van Essen 1991). This diagram and the detailed information about laminar patterns of connections that went alongside it shed new light on the hierarchical organization and division of labor in visual cortex. In the years since, we have seen an almost equally detailed picture of the functional organization of human visual cortex emerge from fMRI studies (Wandell et al. 2007). The significance of having these areas mapped out in humans is that it enables a more direct connection to perception, since one can tie the amount of activity in a given brain area to variations in both stimulus space and psychophysical performance (Heeger 1999; Grill-Spector et al. 2000; Ress and Heeger 2003). This has made it possible to identify areas involved in the representation of three-dimensional form, such as the lateral occipital complex (Kourtzi and Kanwisher 2001). It has also enabled us for the first time to see evidence of "explaining away," in which top-down signals originating from high-level areas appear to decrease the activity in lower level areas when subjects perceive an entire 3D object or scene layout as opposed to its individual parts (Murray et al. 2002).

Some visual areas and neurons exhibit a striking degree of specificity, such as those responsive to faces. Tsao and Livingston used fMRI to localize areas in macaque cortex that are selectively activated by faces and then subsequently recorded in those areas with microelectrodes to characterize responses of individual neurons (Tsao et al. 2006). These studies have revealed a complex of areas that appear to specialize for different aspects of faces such as identity vs. pose (Freiwald et al. 2009). There is now evidence for corresponding areal specializations in humans (Tsao et al. 2008). In addition, Izhak Fried's recordings from the medial temporal lobes in humans have revealed neurons that appear every bit as selective as "grandmother cells," an idea which for years was the subject of theoretical speculation but usually regarded with great skepticism (Quiroga et al. 2005).

Another method that is providing new insights about cortical organization in humans is *neural decoding*. In contrast to traditional approaches that attempt to characterize which class of stimuli a neuron or cortical region responds to, here the goal is to find out what those neurons tell you about the stimulus. When applied to BOLD signals measured over a wide swath of human visual cortex in response to natural images, one finds that lower level areas do a reasonable job at reconstructing image properties such as color and texture, whereas higher level areas reconstruct information about the semantic content of the scene (Naselaris et al. 2009, 2011; Nishimoto et al. 2011). While these particular findings are not surprising given our current understanding of visual cortex, they are nevertheless a testament to the rich, multidimensional information provided by fMRI. Rather than testing specific hypotheses about selected regions of interest, this approach treats the entire 3D volume of BOLD signals as a multielectrode recording array and lets the data speak for itself. Importantly, these studies are most informative when the visual system is presented with complex natural scenes or movies, since these stimuli contain the rich, multidimensional forms of information that are most likely to evoke patterns of activity revealing the functional significance of different brain regions.

How to Infer Scene Geometry from Multiple Views

In parallel with these achievements in neuroscience and psychophysics, the field of computer vision has undergone a number of dramatic advances. Chief among these is the ability to infer three-dimensional scene structure from multiple views, termed *multiple-view geometry* (Hartley and Zisserman 2003). This has been enabled in part by the discovery of stable and unique keypoint detectors and invariant feature descriptors which allow for solving the correspondence problem efficiently (Lowe 2004). It is now possible, given an unordered set of images of the same three-dimensional scene taken from different viewpoints, to simultaneously recover a representation of the 3D scene structure as well as the positions in the scene from which the images were taken (Brown and Lowe 2005). This technology has enabled commercial products such as *Photosynth* which assimilate information from the many thousands of photographs stored on repositories such as Flickr into a unified scene model (Snavely et al. 2006).

While many computer vision algorithms are divorced from biology, there has long been a productive interchange of ideas between the fields of computer vision and biological vision. I believe the advances in multiple-view geometry tell us something important about vision, and that they open the door to a new area of investigation in visual neuroscience—namely, how do animals assimilate the many views they obtain of their environment into a unified representation of the 3D scene? The ability to navigate one's surroundings, to remember where food is, and how to get home is fundamental to the survival of nearly all animals. It would seem to demand an allocentric representation of the 3D environment. However, there has been considerable debate among cognitive psychologists as to whether humans or other animals actually build 3D models as opposed to simply storing 2D views. It is often tacitly assumed that storing 2D views is the simpler, cheaper strategy. But from the standpoint of efficient coding it actually makes the most sense to combine the images acquired while moving through the environment into a single 3D representation, since that is the lowest entropy explanation of the incoming data stream. Now the mathematics and algorithms of multiple-view geometry show us that the computations needed to do this are really quite feasible. In fact these algorithms can run in real-time from video camera input (Newcombe and Davison 2010). The challenge for theorists and modelers now is to figure out how these computations can be performed in a more holistic manner (drawing upon all the data rather than just keypoints), how to exploit the continuity in images over time, and in what format 3D scene information should be represented.

Questions Unanswered

There is little doubt that we are closer to understanding how visual systems work than we were 20 years ago. But how much remains to be understood? Here I shall review areas in which there are still gaping holes in our knowledge. As we shall see, the scope of our ignorance is vast. It is not simply a matter of filling in holes here and there; rather we are missing something fundamental.

How Is Sophisticated Vision Possible in Tiny Nervous Systems?

Much effort in neuroscience is expended to understand how neural circuits in the visual cortex of cats and monkeys enable their perceptual abilities. An often unstated assumption behind these studies is that mammalian cortex is uniquely suited for gaining insight into the neural mechanisms of perception. But one must begin questioning this assumption when confronted with the highly sophisticated visual capabilities found in nervous systems that are smaller by several orders of magnitude.

Consider for example the jumping spider (Fig. 12.4). Unlike other spiders that use a web to extend their sensory space, this animal relies entirely upon vision to localize prey, identify potential mates, and navigate complex terrain. It does so

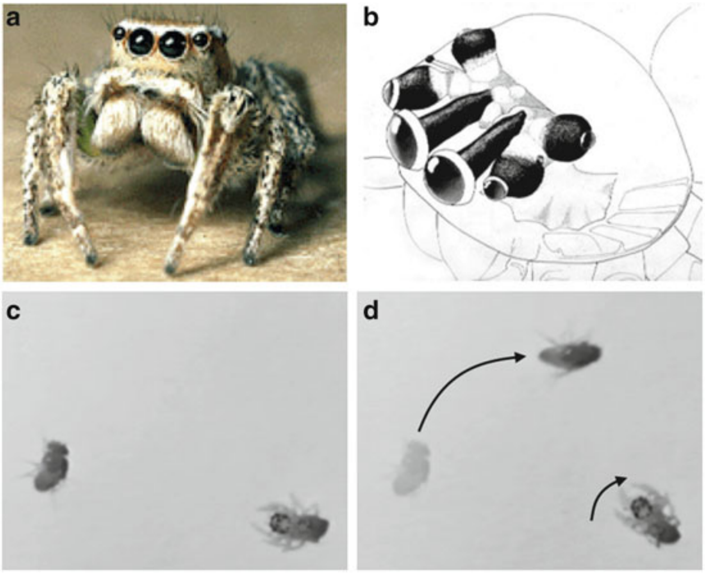

Fig. 12.4 (**a**) Jumping spider (*Habronattus*). (**b**) Jumping spider visual system showing antero-median, antero-lateral, and posterior-lateral eyes. (**c**, **d**) Orienting behavior of a 1-day-old jumping spider (*lower right*) during prey capture. (**a**, **b**) From Wayne Maddison's Tree of Life; (**c**, **d**) video frames filmed by Bruno Olshausen and Wyeth Bair in the Bower lab (Caltech 1991)

using a highly elaborate visual system comprising four pairs of eyes: one pair of frontal facing principal eyes (antero-median eyes) provide a high-resolution image over a narrow field of view, while the other three pairs provide lower resolution images over a wide field of view and are mounted on different parts of the head so as to provide 360° coverage of the entire visual field (Land 1985). Interestingly, the retinae of the antero-median eyes are highly elongated in the vertical direction so as to essentially form a one-dimensional array of photoreceptors. These retinae move from side to side within the head in a smooth (approximately 1 Hz) scanning motion to perform pattern analysis (Land 1969). The jumping spider uses its low resolution system to detect targets or objects of interest, and then orients its body to position the target within the field of view of the high-resolution antero-median eyes for more detailed spatial analysis via scanning (Land 1971).

The jumping spider exhibits a number of striking visual behaviors. Figure 12.4c, d illustrates the tracking and pursuit behavior involved in hunting. The spider initially follows the target (in this case, a fruit fly) with its eye and head movements. It then stalks the fly in a crouching motion before pouncing on it. Mediating this behavior demands the ability to maintain attention on a target, to track the target via appropriate motor commands, and to perform distance estimation. In this case the spider happens to be only 1 day old, so these abilities are largely innate. Another striking visual behavior of the jumping spider is exhibited during courtship, in which the male performs an elaborate dance for the female. During these dances the female visually inspects and attends to the male. Complex pattern recognition via scanning is utilized by both parties during this interaction. Courtship dances may be elicited by presenting a video image of a female (Clark and Uetz 1990), or even a line drawing depicting a jumping spider, to the male (Drees 1952), which further testifies to the role of vision in mediating this behavior. Vision also plays an important role in 3D path planning and navigation. One particular species, *Portia fimbriata*, appears to use its visual system to survey the 3D visual environment before embarking on a path that requires a complex detour to obtain a prey item beyond jumping range (Tarsitano and Jackson 1997; Tarsitano and Andrew 1999).

Thus it would seem that the jumping spider performs complex pattern recognition, visual attention, motion analysis and tracking, distance estimation via stereopsis, and 3D path planning. These are all abilities that most would consider the hallmark of visual cortical function, yet in the jumping spider they are being carried out by a visual system that is no larger than a single hypercolumn of V1, and requiring little or no visual experience during development. There seems to be a huge explanatory gap here between our conventional wisdom and reality.

Another small animal that challenges our conventional wisdom is the sand wasp, *Philanthus triangulum*. The navigational abilities of this animal were intensely studied and described by Tinbergen (1974). He demonstrated that the wasp finds its nest, consisting of a small burrow in the sand, by memorizing the spatial arrangement of debris that happen to immediately surround the nest such as twigs, rocks, or other items. If these items are displaced by a meter or so while the wasp is away hunting, keeping the relative spatial positions of the items intact, it returns to a point in the center of this new arrangement rather than the actual location of its nest. Initially stunned, the animal eventually finds its nest. However, when it next emerges to go out hunting it makes an extra set of circular flights over its nest, as though recommitting to memory the pattern of landmarks surrounding the nest. What is perhaps most astonishing here is that the sand wasp does all of this utilizing only a compound eye, which has very low spatial-resolution. Thus, the complex spatial layout of the environment must somehow be accumulated over time from the dynamic pattern of activity coming from the ommatidia during flight.

It is often tempting to explain away these abilities as the result of simple but clever tricks. To those who try I challenge them to prove such strategies are actually viable by building an autonomous system by these rules that exhibits the same degree of robust, visually guided behavior. Such systems do not exist, and I contend they are still far away from being realized because *we do not understand the fundamental principles governing robust, autonomous behavior in complex environments*. Evolution has discovered these principles and they are embodied in the nervous systems of insects and spiders. There are valuable lessons to be learned from studying them.

The fact that sophisticated visual abilities are present in simpler animals also raises a disturbing question: *If so much can be done with a tiny brain, what more can be done with a large brain?* Perhaps the vast cortical circuits of mammals are carrying out a more complex set of functions than we are currently considering. Perhaps we lack the intellectual maturity needed to ask the right questions about what cortex is doing.

I do not suggest that we must fully understand invertebrate vision as a prerequisite to studying vision in mammals. But I do think that our field is guilty of taking a cortico-centric approach, and that simpler animals have been prematurely dismissed and unjustly neglected in the quest to understand intelligent behavior. One often hears the argument that invertebrates are likely to utilize highly specialized or idiosyncratic neural processing strategies that will not generalize to mammals. But biology is teeming with examples of molecular and cellular mechanisms that are recapitulated across the animal kingdom. Those who study fly genetics are not just interested in flies, they want to know how genes work. At this point there are astonishingly few examples of computations in the nervous system that anyone truly understands. Thus, gaining a solid understanding of neural computation as it occurs in *any* animal would give us much needed insight into the space of possible solutions.

How Do Cortical Microcircuits Contribute to Vision?

Not long after the discovery of orientation selectivity and columnar structure in visual cortex, the view began to emerge that V1 operates as a filter bank in which the image is analyzed in terms of oriented features at different spatial scales (Blakemore and Campbell 1969; De Valois et al. 1982), now often modeled with Gabor functions (Marcelja 1980; Daugman 1985). Others further elaborated on this idea by building hierarchical models composed of successive stages of feature detection and spatial pooling (Fukushima 1980), inspired by Hubel and Wiesel's early proposals (Hubel and Wiesel 1962, 1965). In the ensuing decades, this conceptual framework has come to dominate the theoretical landscape. It has had a profound impact in shaping how neuroscientists form and test hypotheses regarding visual cortical function, and it has influenced the development of computer vision algorithms. It is even referred to as the "standard model" (Riesenhuber and Poggio 2004), and theories that strongly deviate from this framework are often dismissed as biologically implausible. However, this view begins to clash with reality as one takes a closer look at the detailed structure of cortical circuits.

As all students of neuroanatomy know, mammalian neocortex is a layered structure. By convention it has been subdivided into six laminar zones according to various histological criteria such as cell density and morphology. Underlying this overt structure is a detailed microcircuit that connects neurons in a specific way according to the layer they reside in (Douglas et al. 1989; Thomson and Bannister 2003; Douglas and Martin 2004). Inputs from thalamus terminate principally on neurons in layer 4. These neurons in turn project to neurons in layers 2 and 3, which then project back down to layers 5 and 6. Neurons within each layer are recurrently connected by horizontal fibers, with the most extensive of these networks found in layers 2 and 3. Inhibitory interneurons have their own specialized cell types and circuits, and some are interconnected by gap junctions and exhibit synchronous, high gamma oscillations (Mancilla et al. 2007). Layer 1 is mostly composed of the distal tufts of pyramidal cell apical dendrites and the axonal fibers of neurons in other layers. On top of all this, we are beginning to appreciate the "deep molecular diversity" of cortical synapses, which increases the potential complexity of synaptic transmission and plasticity (O'Rourke et al. 2012).

To those who subscribe to the Gabor filter model of V1 I ask, where are these filters? In which layers do they reside, and why do you need such a complex circuit to assemble them? In 1 mm^2 of macaque V1 there are 100,000 neurons, yet the number of LGN afferents innervating this same amount of cortex amounts to the equivalent of only a 14×14 sample node array within the retinal image (Van Essen and Anderson 1995). Why so many neurons for such a small patch of image? To complicate matters further, each neuron is a highly nonlinear device with inputs combining in a multiplicative or "and-like" manner within local compartments of the dendritic tree (Poirazi et al. 2003; Polsky et al. 2004). Such nonlinearities are notably absent from the L-N cascade models commonly utilized within the neural coding community. What are the consequences of these nonlinearities when large numbers of such devices are densely interconnected with one another in a recurrent circuit? It is well known that recurrent networks composed of perceptron-type neurons (linear sum followed by point-wise nonlinearity) can have attractor dynamics, but what are the consequences of dendritic nonlinearities? Is such complexity compatible with the simple notion of a filter or a receptive field? Moreover, why have different layers of processing, and how do the computations and formatting of visual information differ between these layers?

There are numerous hand-wavy explanations and ad hoc models that can be (and have been) constructed to account for all of these things. At the end of the day we are faced with this simple truth: *No one has yet spelled out a detailed model of V1 that incorporates its true biophysical complexity and exploits this complexity to process visual information in a meaningful or useful way*. The problem is not just that we lack the proper data, but that we don't even have the right conceptual framework for thinking about what is happening.

In light of the strong nonlinearities and other complexities of neocortical circuits, one should view the existing evidence for filters or other simple forms of feature extraction in V1 with great skepticism. The vast majority of experiments that claim to measure and characterize "receptive fields" were conducted assuming a linear systems identification framework. We are now discovering that for many V1 neurons these receptive field models perform poorly in predicting responses to complex, time-varying natural images (David et al. 2004; Frégnac et al. 2005; Khosrowshahi et al. 2007). Some argue that with the right amount of tweaking and by including proper gain control mechanisms and other forms of contextual modulation that you can get these models to work (Carandini et al. 2005; Rust and Movshon 2005). My own view is that the standard model is not just in need of revision, *it is the wrong starting point and needs to be discarded altogether.* What is needed in its place is a model that embraces the true biophysical complexity and structure of cortical microcircuits, especially dendritic nonlinearities. The ultimate test of such a model will be in how well it accounts for neural population activity in response to dynamic natural scenes (as opposed to simple test stimuli), and the extent to which it can begin to account for our robust perceptual abilities.

How Does Feedback Contribute to Vision?

At nearly every stage of processing in the visual system, one finds feedback loops in which information flows from one set of neurons to another and then back again. At the very first stage, photoreceptors provide input to a network of horizontal cells which in turn provide negative feedback onto photoreceptors. Hence a photoreceptor does not report a veridical measurement of the amount of light falling upon it, but rather a signal that is modified by context. At later stages, LGN relay neurons provide input to the reticular nucleus which in turn provides negative feedback to LGN relay neurons; LGN projects to V1 and V1 projects back to LGN; V1 projects to V2 which projects back to V1, and so on. What are these feedback loops doing and how do they help us see?

In some cases, such as horizontal cells in the retina, we have fairly good models to suggest what feedback is doing and what it might be good for (i.e., mediating lateral inhibition among photoreceptors to reduce redundancy and increase dynamic range). But in other cases, such as in the thalamo-cortical loop or cortico-cortical loops, there has yet to emerge a clear conceptual model, supported by the data, that tells us what function is being served. There have been numerous experimental attempts to uncover what feedback is doing, for example, by cooling or disabling the neurons in a higher area that feedback onto a lower area and characterizing how response properties in the lower area change (Hupé et al. 2001; Angelucci and Bullier 2003; Andolina et al. 2007). One sees a variety of modulatory effects, but so far there has not emerged a clear consensus or framework for how to incorporate these findings into a larger theory. Indeed there is considerable doubt among neuroscientists as to whether feedback plays any role in dynamically shaping information processing (Lennie 1998).

Perhaps the most striking sign of our conceptual ignorance here is the fact that modern computer vision systems are still largely based on feedforward processing pipelines: image data is preprocessed, features are extracted and then pooled and fed to another layer of processing, or histogrammed and fed to a classifier. One does not typically see algorithms that use the outputs of a higher stage of processing to modify the input coming from a lower stage (though see Arathorn 2005 for a notable exception). In other areas of engineering, such as in the design of control systems or electronic amplifiers, the advantages of feedback are well understood and it is exploited to build robust, stable systems that work in practice. But currently, other than automatic gain control or other early forms of preprocessing, researchers have not discovered how to exploit feedback for more advanced forms of processing that support recognition or other perceptual tasks.

One rationale that is offered in support of feedforward models is that visual recognition occurs so exceedingly fast that there is little time for the iterative type of processing that feedback loops would entail (Thorpe and Imbert 1989). EEG signals correlated with visual recognition in humans arise 150 ms after stimulus onset (Thorpe et al. 1996). In macaque monkey cortex, the earliest neural signals in inferotemporal cortical areas that are discriminative for objects occur ca. 125 ms after stimulus onset (Oram and Perrett 1992; Hung et al. 2005). Given the number of stages of processing and axonal and synaptic delays, it is argued, there is precious little time for any feedback loops to play a significant role in supporting these signals. But this reasoning is based upon overly simplistic and dour assumptions about how feedback works. The conduction velocities of feedforward and feedback axons between V1 and V2 are on the order of 2–4 ms (Angelucci and Bullier 2003). Even between thalamus and V1 the round trip travel time can be as short as 9 ms (Briggs and Usrey 2007). Most importantly though, vision does not work in terms of static snapshots but rather as a dynamical system operating on a continuous, time-varying input stream. Axonal and synaptic delays simply mean that sensory information arriving at the present moment is processed in the context of past information that has gone through a higher level of processing.

Given the space and resource constraints faced by the brain, it seems unlikely that such vast amounts of white matter would be devoted to feedback pathways unless they were serving a useful purpose in shaping information processing. Over the past decade two promising theoretical ideas have been advanced. One is based on the idea of *predictive coding*, in which higher levels send their predictions to lower levels where they are compared, and the residual or degree of mismatch is sent forward (Rao and Ballard 1999). Such a coding scheme would be useful to reduce redundancy and detect novelty. The other is based on *perceptual inference* (or Bayesian inference, as described above) (Lee and Mumford 2003). Here, higher levels also send their predictions to lower levels, but rather than computing differences, the parts where the predictions agree are amplified and the parts where they disagree are suppressed. This type of processing is most useful when lower levels of representation are ambiguous (such as the aperture problem in the computation of motion). Higher level knowledge and context are used to adjudicate between different interpretations and resolve ambiguity. Formally this may be cast in terms of probabilistic inference in graphical models or "belief propagation." To validate either of these hypotheses one would need to investigate the effects of feedback during the viewing of natural images or other complex, structured images where prediction can play a role, or the need for disambiguation arises. Indeed this may explain why the findings of previous experiments using simplified test stimuli have been rather inconclusive.

What Is the Role of Neuronal Oscillations in Visual Processing?

Since Hans Berger's first EEG measurements in the 1920s it has been known that the brain oscillates. Early investigators ascribed the terms *alpha*, *beta*, and *gamma* to oscillations occurring in different frequency bands, and they attempted to relate these oscillations to various states of arousal, perception, cognition, or clinical pathologies. Later, when neurophysiologists such as Barlow, Kuffler, Hubel, and Wiesel began achieving success with single-unit recordings, attention turned to the activity of individual neurons. Interest in oscillations dissipated, and the focus instead shifted to studying how the *stimulus-driven* firing rate of neurons encodes features of the visual world. Against this backdrop in 1989, Gray and Singer showed that the activity of single neurons in V1 is phase-locked to gamma oscillations in the local field potential (LFP), and furthermore that the degree of synchrony between neurons depends on whether the features they encode belong to a common object (Gray and Singer 1989). This finding reignited interest in oscillations, especially among theorists who speculated that they may serve as a mechanism for feature binding and attention, or even consciousness. Experimentalists argued among themselves as to whether oscillations or synchrony were actually present. Sides were taken and debates were staged (e.g., at the Society for Neuroscience 1993 annual meeting), and each side argued passionately for their point of view.

Now almost 20 years later the debate has mostly subsided. Few doubt the existence of oscillations—they have withstood the test of time and have been shown to be a ubiquitous property of sensory systems, from the locust olfactory system to the mammalian retina and visual cortex. One senses that the field has settled into taking a more dispassionate approach to investigate what causes these oscillations, under

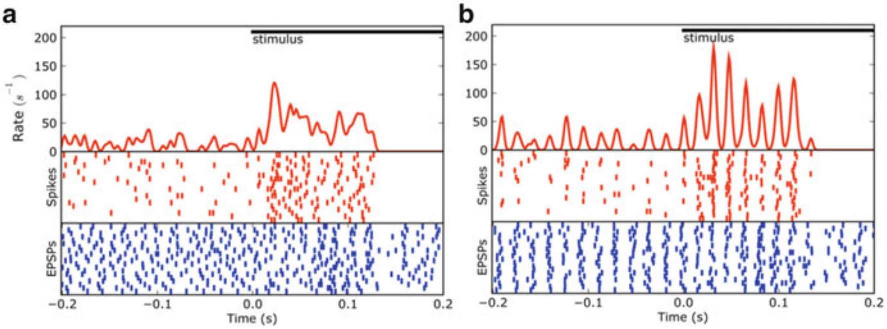

Fig. 12.5 LGN neurons synchronize to 50 Hz retinal oscillations. (**a**) PSTH and spike rasters in response to repeated presentations of a stimulus. Note the apparent variability in the latency of the LGN neuron's response. (**b**) When the LGN spikes are realigned to the instantaneous phase of retinal oscillations extracted from the EPSPs for each trial, the variability in response latency is vastly reduced (from Koepsell et al. 2009)

what conditions they arise, and how they relate to perception. However, there is still little concrete evidence that suggests what they are doing and how they help us see.

One recent finding that I believe points to an important role for oscillations in vision comes from recordings from cat LGN neurons in Judith Hirsch's laboratory (Koepsell et al. 2009). These data reveal that the spiking activity of some neurons in the LGN is phase-locked to the 50 Hz oscillations arising from the retina. These oscillations are readily apparent in the electro-retinogram and have been observed in recordings from retinal neurons, but their effect on downstream processing was previously unknown. Koepsell et al. showed that when the phase of these ongoing oscillations is taken into account, the apparent variability in the response latency of LGN neurons—commonly attributed to "noise"—is vastly reduced (Fig. 12.5). In other words, LGN neurons exhibit a much higher degree of temporal precision—and hence information carrying capacity—when the phase of ongoing oscillations is included in reading out their activity (as opposed to considering the stimulus-driven component only). What could this extra information be used for? Koepsell and Sommer propose that oscillations propagating through distributed networks in the retina could be used to compute "graph cuts," an effective method of image segmentation that is widely used in computer vision (Koepsell et al. 2010). In their model, the firing rate of a neuron encodes contrast and the phase of oscillation encodes region membership. While highly speculative, the theory nevertheless demonstrates how oscillations could be leveraged in a profound and elegant way to carry out computations requiring the rapid and global spread of information across an image to solve a difficult problem in vision.

When considering oscillation-based theories it is important to bear in mind that the prevailing rate-based, stimulus-driven view of neural function, while often portrayed as fact, is itself a theory. Though there are countless examples where firing rate correlates with perceptual variables, this in itself does not demonstrate that information is actually encoded and read out this way. So little is known at this point that there is much room for alternative theories. But if one accepts that neural activity is an information-bearing signal in the brain, then oscillations and other forms of ongoing activity must be included in a full account of neural function.

How to Build Robust, Autonomous Vision Systems?

In 1973 Sir James Lighthill issued a report to the British Parliament that condemned AI for failing to achieve its grandiose objectives and recommended that its funding be cut off (the recommendation was subsequently adopted, killing AI research in the UK for nearly a decade). At the center of his argument was that robotic systems were only capable of operating in restricted domains, and that scaling up to general purpose intelligence that could deal with real world conditions would require a combinatorial explosion in computational resources. The idea that we might someday build general purpose robots, he claimed, was a “mirage.” A debate was held between Lighthill and three leading AI researchers, Donald Michie, John McCarthy, and Richard Gregory, who defended their aims and work as realistic and worthwhile (BBC 1973). State-of-the-art robots of the day such as SRI’s *Shakey* and Edinburgh’s *Freddy* took center stage to illustrate the promising achievements of AI. These robots could perceive the world through cameras that extracted the outlines of objects and could guide an actuator to grasp or manipulate the objects. They could execute complex tasks, such as assembling a toy car from parts randomly arranged on a table, in a completely autonomous manner without human intervention.

Now almost 40 years later, with all of the participants of that debate gone, it is almost too painful to ask this, but … was Lighthill right? Consider that over this span of time Moore’s law has brought us an increase of *six orders of magnitude* in available computational resources. Can we claim that robots have similarly advanced compared to their predecessors in the early 1970s? *Stanley*, the robot that won DARPA’s Grand Challenge desert road race in 2005, is heralded as a triumph of AI. But upon closer examination it would seem to exemplify exactly the sort of domain-specific limitations that Lighthill railed against—it was preprogrammed with a map of the entire route and 3000 GPS waypoints, and it followed a road with few major obstacles on a bright sunny day. As such, it was primarily a test of high-speed road finding, obstacle detection, and avoidance in desert terrain (Thrun et al. 2006). Its success in navigating the course was mainly the result of clever engineering—Kalman filters to compute robust, optimal estimates of position, and combining LIDAR and image data to find drivable terrain and stay in the center of the road. These are notable achievements, but it is difficult to imagine that this is the level of visual intelligence that Michie, McCarthy, and Gregory would have hoped to see emerge by the early twenty-first century.

Now consider these robots in comparison to the jumping spider or sand wasp. To survive they must navigate unfamiliar, complex terrain that is filled with obstacles, variable illumination from shadows, and potentially unstable surfaces. They have no GPS way points or roads to provide guidance. Rather, they must acquire and store

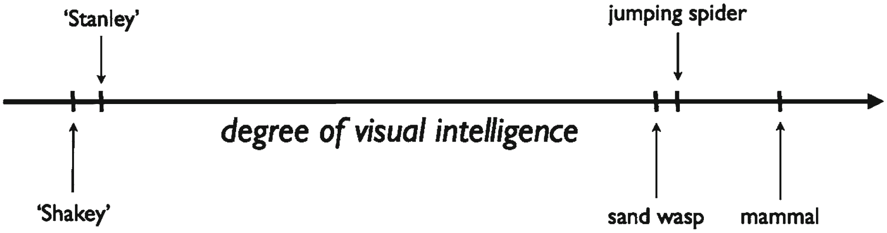

Fig. 12.6 When measured in terms of visual intelligence, there is still a wide gulf separating robots such as Shakey and Stanley from biological visual systems

information about the environment as they go so as to remember where they have been, where the food is, and how to get home. They must detect, localize, track, and successfully capture prey, even when seen against complex backgrounds. They must deal with unforeseen events such as getting knocked off course by wind or debris. They must continue to function 24/7 in the face of the elements such as rain or dust or changes in lighting conditions. And they do all of this while consuming only minuscule amounts of power in comparison to their robotic counterparts.

While *Stanley* unquestionably represents an advance over *Shakey*, both of these systems would seem equally far removed from the jumping spider or sand wasp, let alone humans, when measured in terms of the level of robust, autonomous behavior they exhibit (Fig. 12.6). That we stand at this impasse after 40 years I believe tells us something important. It suggests that the problem we face is not just technological but rather due to a scientific gap in our knowledge. *We are missing something fundamental about the principles of vision and how it enables autonomous behavior.* Computing optic flow or building a depth map of a scene, while useful, is not sufficient to robustly navigate, interact with, and survive in the natural three-dimensional environment. What exactly *is* needed is of course difficult to say—that is the problem we are up against. But I would point to two things. One is a richer representation of surface layout in the surrounding environment that expresses not only its 3D geometry but also its *affordances*—that is, the actions that are possible (Gibson 1986). The other is to move beyond the Turing machine, procedural framework that today's robots are trapped in—that is, an infinite loop of "acquire data," "make decisions," and "execute actions." What is needed is a more fluid, dynamic interaction between perception and action. Theories for how to do this are now beginning to emerge but it is a field still in its infancy (Gordon et al. 2011).

Questions Not Yet Asked

The answers we get from experiments are only as useful as the questions we ask. The key is to ask the right questions to begin with. But how do we know what these are? Most of the questions described in the preceding section are ones that scientists are already keenly aware of and which drive current research efforts. Undoubtedly though there are other important questions that no one working in the field today has even thought to ask yet, just as computer vision researchers in the 1960s never thought to ask how you find the edges of an object in a scene. This points to the importance of another process of discovery beyond answering questions—that is, discovering the questions that need to be asked.

Here I will suggest two ways that we can accelerate the process of discovering what these questions are. One is to take an *exploratory approach* that casts a wide net and seeks to reveal interesting phenomena. The other is to educate ourselves about the problems of vision by attempting to *build* neuromorphic visual systems that enable autonomous behavior.

The Need for Exploratory Approaches

Scientists by their nature are eager to test hypotheses or to tell a story about how a given set of facts or findings fit together and explain perceptual phenomena. But as we have seen, vision presents us with deep computational problems, and nervous systems confront us with stunning complexity. Most of the hypotheses we test and the stories we tell are far too simple minded by comparison, and ultimately they turn out to be wrong. Worse yet, they can be misleading and stifling because they encourage one to look at the data through a narrow lens. When one carefully designs a set of experiments to test a specific set of hypotheses, the data obtained are often of little value for looking at other issues. In some cases this may be warranted, but when the hypothesis landscape is not well formed to begin with it may be more worthwhile to take an exploratory approach.

The exploratory approach is more observational in nature. The goal is to document how the system works in its natural state—for example, what are the distributions of firing rates among neurons in different layers, and in different cortical areas, during natural vision? Such experiments do not test any particular hypothesis, and the outcome may simply be a large table of numbers. But such data would be of immense value in helping us to understand what kind of a system we are dealing with, and they are of pivotal importance in shaping theories.

Another goal of the exploratory approach is discover new phenomena that surprise us and defy conventional wisdom. These can then provide clues about what we *should* be looking for. A notable example is the discovery of orientation selectivity. The idea that visual neurons might be selective to lines or edges at different orientations did not occur to Hubel and Wiesel a priori. Rather, they were probing the visual cortex with spots of light using a slide projector, and in the process of moving slides in and out of the projector they noticed that the edge of the slide moving over the receptive field happened to elicit a robust neural response (Hubel 1982). This observation in turn led to a revolution in visual neuroscience. Tinkering is often frowned upon in scientific circles, especially by study sections and review panels of the major scientific funding bodies. But when one is mostly in the dark to begin with—as I would argue we are in our understanding of the visual cortex—a certain amount of tinkering seems warranted.

I do not advocate that we abandon the hypothesis-based approach—it has formed the bedrock of modern science because in many cases it has been a fruitful and productive path to knowledge. But we should recognize when this approach is appropriate and when it is not. Storytelling makes science interesting, and it often makes a finding seem more compelling, but it can also lead to a false sense of complacency, a feeling that we have understood something when in fact the real story is orders of magnitude more complicated. We should be more inclined to take these stories with a grain of salt and instead be on the lookout for something deeper lurking beneath the surface. And no one should feel ashamed to report a complete, unfiltered set of findings without a story to envelop them. After all, one person's untidy finding may provide the missing piece in another person's theory.

Learning About Vision by Building Autonomous Systems

There is very little that neuroscience per se has taught us about the principles of vision. That we know there is a ventral and dorsal stream, a hierarchy of visual areas, and neurons that selectively respond to certain visual features in these areas does not tell us *what* problems are being solved and *how*. They provide strong hints and tantalizing clues to be sure, but trying to build a functional vision system by directly mimicking these attributes in a computer chip is like trying to build a flying machine out of flapping wings and feathers.

By contrast, the failures of robot vision in the 1960s were a transformative learning experience in the study of vision. They set the stage for people like David Marr to intensely study the computational problems of vision and to theorize how biological vision systems work. The field thus made an advance by trying to solve an important and unsolved problem, the depth of which was previously unappreciated. I believe this will continue to be the case in the future—we will learn the most about the principles of vision by attempting to build autonomous vision systems, learning what works and what does not, and then drawing upon these insights in studying the visual systems of humans and other animals.

To some extent this is a role that computer vision already plays. However, mainstream computer vision is focused on solving a prescribed set of problems that have been defined by computer scientists and engineers. Algorithms for shape from shading, optic flow, and stereo are judged by how well they perform on standard benchmarks, where the correct representation is assumed to be known. Object recognition is distilled down to a problem of classification, one of converting pixels to labels, again with benchmark datasets for judging performance. If we wish to gain insight into the principles of biological vision, or autonomous visual behavior in general, it will require a different approach.

What is needed is an approach that, like computer vision, attempts to solve problems, but where more attention is paid to how we define those problems, and the computational architectures we draw upon to solve them. The choice of problems should be guided by animal behavior and psychophysics: What are the tasks that animals need to solve in order to survive in the natural environment? What are the performance characteristics of human or other animal observers in these tasks? In addition, it is important to take into account and exploit the unique computational properties of neural systems, what Carver Mead called "neuromorphic engineering." The only functional vision systems we know of today are built out of nonlinear recurrent networks, they compute with analog values, and they run in continuous time. They are not Turing machines. Thus, in considering the space of solutions to visual problems this needs to be taken into account.

Finally, it is important to bear in mind that vision did not evolve as a stand-alone function, but rather as part of the perception–action cycle. As philosopher Robert Cummins put it, "Why don't plants have eyes?" We have much to gain by building vision systems with tight sensorimotor loops and learning what problems need to be overcome in doing so. This area remains vastly under investigated, and is likely to uncover to many questions that have yet to be asked.

Acknowledgments I thank Jim Bower for encouraging me to write this article and for his patience in giving me the time to complete it, and Jim DiCarlo for providing the MIT AI Memo. Supported by funding from NSF (IIS-1111654), NIH (EY019965), NGA (HM1582-08-1-0007), and the Canadian Institute for Advanced Research.

References

Adelson EH (1993) Perceptual organization and the judgment of brightness. Science 262(5142):2042–2044

Adelson EH (2000) Lightness perception and lightness illusions. In: Gazzaniga M (ed) The new cognitive neurosciences, 2nd edn. MIT, Cambridge, MA, pp 339–351

Andolina IM, Jones HE et al (2007) Corticothalamic feedback enhances stimulus response precision in the visual system. Proc Natl Acad Sci USA 104(5):1685–1690

Angelucci A, Bullier J (2003) Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? J Physiol Paris 97(2–3):141–154

Arathorn DW (2005) Computation in the higher visual cortices: map-seeking circuit theory and application to machine vision. In: Proceedings of the 33rd applied imagery pattern recognition workshop (AIPR 2004), Washington, DC, 1–6

Arathorn DW, Yang Q et al (2007) Retinally stabilized cone-targeted stimulus delivery. Opt Express 15(21):13731–13744

Atick J, Redlich A (1992) What does the retina know about natural scenes? Neural Comput 4:196–210

Barlow H (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA (ed) Sensory communications. MIT, Cambridge, MA, pp 217–234

Barron JT, Malik J (2012) Shape, albedo, and illumination from a single image of an unknown object. In: Conference on computer vision and pattern recognition, Washington, DC, 1–8

BBC (1973) Controversy. http://www.aiai.ed.ac.uk/events/lighthill1973/1973-BBC-Lighthill-Controversy.mov

Blakemore C, Campbell FW (1969) On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. J Physiol 203(1):237–260

Boyaci H, Fang F, Murray SO, Kersten D (2007) Responses to lightness variations in early human visual cortex. Curr Biol 17:989–993

Brainard DH, Williams DR et al (2008) Trichromatic reconstruction from the interleaved cone mosaic: Bayesian model and the color appearance of small spots. J Vis 8(5):15

Briggs F, Usrey WM (2007) A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey. J Neurosci 27(20):5431–5436

Brown M, Lowe DG (2005) Unsupervised 3D object recognition and reconstruction in unordered datasets. In: Fifth international conference on 3-D digital imaging and modeling, 2005, 3DIM 2005, Ottawa, 56–63

Cadieu CF, Olshausen BA (2012) Learning intermediate-level representations of form and motion from natural movies. Neural Comput 24(4):827–866

Carandini M, Demb JB et al (2005) Do we know what the early visual system does? J Neurosci 25(46):10577–10597

Clark D, Uetz G (1990) Video image recognition by the jumping spider, Maevia inclemens(Araneae: Salticidae). Anim Behav 40:884–890

Dan Y, Atick JJ et al (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J Neurosci 16(10):3351–3362

Daugman J (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am 2:1160–1169

David SV, Vinje WE et al (2004) Natural stimulus statistics alter the receptive field structure of v1 neurons. J Neurosci 24(31):6991–7006

Dayan P, Hinton GE et al (1995) The Helmholtz machine. Neural Comput 7(5):889–904

De Valois RL, Albrecht DG et al (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res 22(5):545–559

Douglas RJ, Martin KAC (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27(1):419–451

Douglas RJ, Martin KAC et al (1989) A canonical microcircuit for neocortex. Neural Comput 1(4):480–488

Drees O (1952) Untersuchungen über die angeborenen verhaltensweisen bei springspinnen (salticidae). Z Tierpsychol 9(2):169–207

Dreyfus HL, Dreyfus SE (1988) Making a mind versus modeling the brain: artificial intelligence back at a branchpoint. Daedalus 117(1):15–43

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1(1):1–47

Field D (1987) Relations between the statistics of natural images and the response properties of cortical-cells. J Opt Soc Am A 4:2379–2394

Frégnac Y, Baudot P, Levy M, Marre O (2005) An intracellular view of time coding and sparseness in V1 during virtual oculomotor exploration of natural scenes. In: 2nd International Cosyne conference in computational and systems neuroscience, Salt Lake City, UT, 17

Freiwald WA, Tsao DY et al (2009) A face feature space in the macaque temporal lobe. Nat Neurosci 12(9):1187–1196

Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36(4):193–202

Gauthier JL, Field GD et al (2009a) Receptive fields in primate retina are coordinated to sample visual space more uniformly. PLoS Biol 7(4):e1000063

Gauthier JL, Field GD et al (2009b) Uniform signal redundancy of parasol and midget ganglion cells in primate retina. J Neurosci 29(14):4675–4680

Geisler WS, Perry JS et al (2001) Edge co-occurrence in natural images predicts contour grouping performance. Vision Res 41(6):711–724

Gibson J (1986) The ecological approach to visual perception—James Jerome Gibson—Google books. Erlbaum, Hillsdale, NJ

Gordon G, Kaplan DM, Lankow B, Little DY, Sherwin J, Suter BA, Thaler L (2011) Toward an integrated approach to perception and action: conference report and future directions. Front Syst Neurosci 5:20

Gray CM, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci USA 86:1698–1702

Grill-Spector K, Kushnir T et al (2000) The dynamics of object-selective activation correlate with recognition performance in humans. Nat Neurosci 3(8):837–843

Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press, Cambridge, MA

Heeger DJ (1999) Linking visual perception with human brain activity. Curr Opin Neurobiol 9(4):474–479

Hofer H, Williams D (2005) Different sensations from cones with the same photopigment. J Vis 5(5):444–454

Hofer H, Carroll J et al (2005) Organization of the human trichromatic cone mosaic. J Neurosci 25(42):9669–9679

Hubel DH (1982) Exploration of the primary visual cortex, 1955–78. Nature 299(5883):515–524

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Hubel DH, Wiesel TN (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 28:229–289

Hung CP, Kreiman G et al (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310(5749):863–866

Hupé JM, James AC et al (2001) Feedback connections act on the early part of the responses in monkey visual cortex. J Neurophysiol 85(1):134–145

Hyvarinen A, Hoyer PO (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Res 41(18):2413–2423

Hyvarinen A, Gutmann M et al (2005) Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. BMC Neurosci 6(1):12

Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290(5802):91–97

Karklin Y, Lewicki M (2003) Learning higher-order structures in natural images. Network 14(3):483–499

Karklin Y, Lewicki M (2005) A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. Neural Comput 17(2):397–423

Karklin Y, Lewicki MS (2009) Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457(7225):83–85

Kersten D, Mamassian P et al (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304

Khosrowshahi A, Baker J et al (2007) Predicting responses of V1 neurons to natural movies. Society for Neuroscience, San Diego, CA, p 33

Knill D, Richards W (1996) Perception as Bayesian inference. Cambridge University Press, Cambridge, MA

Knill DC, Saunders JA (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43(24):2539–2558

Koepsell K, Wei Y, Wang Q, Rathbun DL, Usrey WM, Hirsch JA, Sommer FT (2009) Retinal oscillations carry visual information to cortex. Front Syst Neurosci 3

Koepsell K, Wang X et al (2010) Exploring the function of neural oscillations in early sensory systems. Front Neurosci 4:53

Kourtzi Z, Kanwisher N (2001) Representation of perceived object shape by the human lateral occipital complex. Science 293(5534):1506–1509

Land MF (1969) Movements of the retinae of jumping spiders (Salticidae: dendryphantinae) in response to visual stimuli. J Exp Biol 51(2):471–493

Land MF (1971) Orientation by jumping spiders in the absence of visual feedback. J Exp Biol 54(1):119–139

Land MF (1985) Fields of view of the eyes of primitive jumping spiders. J Exp Biol 119:381–384

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20(7):1434–1448

Lee TS, Yang CF et al (2002) Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. Nat Neurosci 5(6):589–597

Lennie P (1998) Single units and visual cortical organization. Perception 27(8):889–935

Litke AM, Bezayiff N et al (2004) What does the eye tell the brain? Development of a system for the large-scale recording of retinal output activity. IEEE Trans Nucl Sci 51(4):1434–1440

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

Ma WJ, Beck JM et al (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9(11):1432–1438

Mamassian P, Knill DC et al (1998) The perception of cast shadows. Trends Cogn Sci 2(8):288–295

Mancilla JG, Lewis TJ et al (2007) Synchronization of electrically coupled pairs of inhibitory interneurons in neocortex. J Neurosci 27(8):2058–2073

Marcelja S (1980) Mathematical description of the responses of simple cortical cells. J Opt Soc Am 70(11):1297–1300

Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. WH Freeman, San Francisco, CA

Mumford D (1994) Neuronal architectures for pattern-theoretic problems. Large-scale neuronal theories of the brain. MIT, Cambridge, MA

Murray SO, Kersten D et al (2002) Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci USA 99(23):15164–15169

Nakayama K, He Z et al (1995) Visual surface representation: a critical link between lower-level and higher-level vision. In: Kosslyn SM, Osherson DN (eds) An invitation to cognitive science: visual cognition, vol 2. MIT, Cambridge, MA, pp 1–70

Naselaris T, Prenger RJ et al (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63(6):902–915

Naselaris T, Kay KN et al (2011) Encoding and decoding in fMRI. Neuroimage 56(2):400–410

Newcombe RA, Davison AJ (2010) Live dense reconstruction with a single moving camera. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR), San Francisco, CA, 1498–1505

Nishimoto S, Vu AT et al (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21(19):1641–1646

O'Rourke NA, Weiler NC, Micheva KD, Smith SJ (2012) Deep molecular diversity of mammalian synapses: why it matters and how to measure it. Nat Rev Neurosci 13:365–379

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583):607–609

Oram MW, Perrett DI (1992) Time course of neural responses discriminating different views of the face and head. J Neurophysiol 68(1):70–84

Papert S (1966) The summer vision project. MIT Artificial Intelligence Group, Vision Memo No. 100, 1–6

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference—Judea Pearl—Google books. Morgan Kaufmann Publishers, San Francisco, CA

Poirazi P, Brannon T et al (2003) Pyramidal neuron as two-layer neural network. Neuron 37(6):989–999

Polsky A, Mel BW et al (2004) Computational subunits in thin dendrites of pyramidal cells. Nat Neurosci 7(6):621–627

Qiu FT, von der Heydt R (2005) Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. Neuron 47(1):155–166

Quiroga RQ, Reddy L et al (2005) Invariant visual representation by single neurons in the human brain. Nature 435(7045):1102–1107

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87

Rao RPN, Olshausen BA et al (2002) Probabilistic models of the brain: perception and neural function. MIT, Cambridge, MA

Ress D, Heeger DJ (2003) Neuronal correlates of perception in early visual cortex. Nat Neurosci 6(4):414–420

Riesenhuber M, Poggio T (2004) How the visual cortex recognizes objects: the tale of the standard model. In: Chalupa L, Werner J (eds) The visual neurosciences. MIT, Cambridge, MA, pp 1–14

Roorda A (2011) Adaptive optics for studying visual function: a comprehensive review. J Vis 11(7)

Roorda A, Williams DR (1999) The arrangement of the three cone classes in the living human eye. Nature 397(6719):520–522

Rossi AF, Rittenhouse CD et al (1996) The representation of brightness in primary visual cortex. Science 273(5278):1104–1107

Rust N, Movshon J (2005) In praise of artifice. Nat Neurosci 8(12):1647–1650

Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4:819–825

Sincich LC, Zhang Y et al (2009) Resolving single cone inputs to visual receptive fields. Nat Neurosci 12(8):967–969

Smith E, Lewicki M (2006) Efficient auditory coding. Nature 439(7079):978–982

Snavely N, Seitz SM et al (2006) Photo tourism: exploring photo collections in 3D. ACM, New York, NY

Srinivasan MV, Laughlin SB et al (1982) Predictive coding: a fresh view of inhibition in the retina. Proc R Soc Lond B Biol Sci 216(1205):427–459

Tappen MF, Freeman WT et al (2005) Recovering intrinsic images from a single image. IEEE Trans Pattern Anal Mach Intell 27(9):1459–1472

Tarsitano M, Andrew R (1999) Scanning and route selection in the jumping spider Portia labiata. Anim Behav 58(2):255–265

Tarsitano M, Jackson RR (1997) Araneophagic jumping spiders discriminate between detour routes that do and do not lead to prey. Anim Behav 53:257–266

Thomson AM, Bannister AP (2003) Interlaminar connections in the neocortex. Cereb Cortex 13(1):5–14

Thorpe SJ, Imbert M (1989) Biological constraints on connectionist models. In: Pfeifer R, Schreter Z, Fogelman-Soulie F, Steels L (eds) Connectionism in perspective. Elsevier Inc., Amsterdam, pp 63–92

Thorpe S, Fize D et al (1996) Speed of processing in the human visual system. Nature 381:520–522

Thrun S, Montemerlo M et al (2006) Stanley: the robot that won the DARPA grand challenge. J Field Robot 23(9):661–692

Tinbergen N (1974) Curious naturalists (revised edition). University of Massachusetts Press, Amherst, MA

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12(1):97–136

Tsao DY, Freiwald WA et al (2006) A cortical region consisting entirely of face-selective cells. Science 311(5761):670–674

Tsao DY, Moeller S et al (2008) Comparing face patch systems in macaques and humans. Proc Natl Acad Sci 105(49):19514–19519

Van Essen DC, Anderson CH (1995) Information processing strategies and pathways in the primate visual system. In: Zornetzer SF, Davis JL, Lau C, McKenna T (eds) An introduction to neural and electronic networks. Academic, San Diego, CA, pp 45–76

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287(5456):1273–1276

Vogel CR, Arathorn DW et al (2006) Retinal motion estimation in adaptive optics scanning laser ophthalmoscopy. Opt Express 14(2):487–497

Wandell BA, Dumoulin SO et al (2007) Visual field maps in human cortex. Neuron 56(2):366–383

Zhou H, Friedman HS et al (2000) Coding of border ownership in monkey visual cortex. J Neurosci 20(17):6594–6611

Zhu M, Rozell C (2011) Population characteristics and interpretations of nCRF effects emerging from sparse coding. In: Computational and Systems Neuroscience (COSYNE), Salt Lake City, UT