Object Categorization: Computer and Human Vision Perspectives

Edited by

Sven Dickinson, Ales Leonardis, Bernt Schiele, and Michael Tarr

1

On what it means to see, and what we can do about it

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853-7601, USA
http://kybele.psych.cornell.edu/∼edelman

Seeing is forgetting the name of the thing one sees.
Paul Valéry (1871-1945)

If you are looking at the object, you need not think of it.[1]
Ludwig Wittgenstein (1889-1951)

1.1 Introduction

A decisive resolution of the problems of high-level vision is at present impeded not by a shortage of computational ideas for processing the array of measurements with which vision begins, but rather by certain tacit assumptions behind the very formulation of these problems.

Consider the problem of object recognition. Intuitively, recognition means determining whether or not the input contains a manifestation of a known object, and perhaps identifying the object in question. This intuition serves well in certain contrived situations, such as character recognition in reading or machine part recognition in an industrial setting — tasks that are characterized first and foremost by only involving objects that come from closed, well-defined sets. An effective computational strategy for object recognition in such situations is to maintain a library of object templates and to match these to the input in a flexible and efficient manner (Basri and Ullman, 1988; Edelman et al., 1990; Huttenlocher and Ullman, 1987; Lowe, 1987).

In categorization, where the focus of the problem shifts from identifying concrete shapes to making sense of shape *concepts*, this strategy begins to unravel — not because flexible template matching as such cannot

keep up with the demands of the task, but rather because the template library is no longer well-defined at the levels of abstraction on which the system must operate. The established approaches to both recognition and categorization are thus seen to suffer from the same shortcoming: an assumption that the input is fully interpretable in terms of a finite set of well-defined visual concepts or "objects."

In this chapter, I argue that forcing a specific and full conceptual interpretation on a given input may be counterproductive not only because it may be a wrong conceptual interpretation, but also because the input may best be left altogether uninterpreted in the traditional sense. Non-conceptual vision is not widely studied, and yet it seems to be the rule rather than the exception among the biological visual systems found on this planet, including human vision in its more intriguing modes of operation (Edelman, 2008, ch.5).

To gain a better understanding of natural vision, and to make progress in designing robust and versatile artificial visual systems, we must therefore start at the beginning, by carefully considering the range of tasks that natural vision has evolved to solve. In other words, we must sooner rather than later face up to the question of what it means to see.

1.2 Seeing vs. "seeing as"

In his epochal book *Vision*, David Marr (1982) offered two answers to the question of what it means to see: one short and intuitive, the other long, detailed, and computational. Briefly, according to Marr, to see means "to know what is where by looking" — a formulation that expresses the computational idea that vision consists of processing images of a scene so as to *make explicit* what needs to be known about it. On this account, "low-level" vision has to do, among other things, with recovering from the stimulus the positions and orientations of visible surfaces (perhaps in the service of navigation or manipulation), and "high-level" vision with determining which of the known objects, if any, are present in the scene.

The research program initiated by Marr and Poggio (1977), now in its fourth decade, spurred progress in understanding biological vision and contributed to the development of better machine vision systems. Most of the progress has, however, been confined to the understanding of vision *qua* interpretation, rather than of vision *per se*. The difference between the two is best introduced with a selection of passages from

Wittgenstein (1958), who distinguished between "seeing" and "seeing as":

> Two uses of the word "see."
> The one: "What do you see there?" — "I see *this*" (and then a description, a drawing, a copy). The other: "I see a likeness between these two faces" [...]
> I contemplate a face, and then suddenly notice its likeness to another. I *see* that it has not changed; and yet I see it differently. I call this experience "noticing an aspect." [...]
> I suddenly see the solution of a puzzle-picture. Before, there were branches there; now there is a human shape. My visual impression has changed and now I recognize that it has not only shape and color but also a quite particular 'organization.' [...]
> Do I really see something different each time, or do I only interpret what I see in a different way? I am inclined to say the former. But why? — To interpret is to think, to do something; seeing is a state.
>
> — Wittgenstein (1958, part II, section xi)

A little reflection reveals that the two kinds of seeing — I'll call the first one "just seeing" to distinguish it from "seeing as" — are related to each other. Informally, the ultimate level of "just seeing" would be attained by a system that can see any possible scene "as" anything at all — that is, a system that can parse differences among scenes in every conceivable way, by varying the labels it attaches to each discernible "aspect" of the input, to use Wittgenstein's expression (these aspects need not be spatial).[2]

Semi-formally, the power of a visual system can be quantified by treating scenes as points in some measurement space, $s \in \mathcal{S}$, which are to be distinguished from one another by being classified with respect to a set of concepts $\mathcal{C}$. A system is powerful to the extent that it has both a high-resolution measurement front end and a sophisticated conceptual back end (a 12-megapixel digital camera and a person with low vision are both not very good at seeing, for complementary reasons). If, however, the dimensionality of the measurement space is sufficiently high, the system in question will be able at least to *represent* a very large variety of distinct scenes.[3] Let us, therefore, assume that the dimensionality of the measurement space is in the mega-pixel range (as indeed it is in the human retina) and proceed to examine the role of conceptual sophistication in seeing.

This can be done by formalizing the visual system's conceptual back end as a classification model. The model's power can then be expressed in terms of its Vapnik-Chervonenkis or VC dimension (Vapnik, 1995;

Vapnik and Chervonenkis, 1971). Consider a class of binary concepts $f \in \mathcal{C}$ defined over a class of inputs (that is, measurements performed over scenes from $\mathcal{S}$), such that $f : \mathcal{S} \rightarrow \{0, 1\}$. The VC dimension $VCdim(\mathcal{C})$ of the class of concepts (that is, of the model that constitutes the categorization back end of the visual system) quantifies its ability to *distinguish* among potentially different inputs. Specifically, the $VCdim$ of a concept class $\mathcal{C}$ is defined as the cardinality of the largest set of inputs that a member concept can shatter.[4]

Because classifying a scene as being an instance of a concept amounts to seeing it *as* something, we have thus effectively formalized the notion of "seeing as." We are now ready to extend this framework to encompass the ability to "just see." The key observation is this: among several conceptual systems that happen to share the same measurement space, the one with the highest VC dimension is the most capable of distinguishing various subtle aspects of a given input. In other words, to progressively more complex or higher-$VCdim$ visual systems, the same scene would appear richer and more detailed — a quality that translates into the intuitive notion of a progressively better ability to "just see."

It is worth recalling that the VC dimension of a class of visual concepts determines its learnability: the larger $VCdim(\mathcal{C})$, the more training examples are needed to reduce the error in generalizing $\mathcal{C}$ to new instances below a given level (Blumer et al., 1986; Edelman, 1993). Because in real-life situations training data are always at a premium (Edelman and Intrator, 2002), and because high-$VCdim$ classifiers are too flexible and are therefore prone to overfitting (Baum and Haussler, 1989; Geman et al., 1992), a purposive visual system should always employ the simplest possible classifier for each task that it faces. For this very reason, purposive systems that are good at learning from specific experiences are likely also to be poor general experiencers: non-conceptual and purposeless experience of "just seeing" means being able to see the world under as many as possible of its different aspects, an ability which corresponds to having a high $VCdim$.[5]

To clarify this notion, let us now imagine some examples. A rather extreme one would be a pedestrian avoidance system installed in a car, which sees any scene s that's in front of it either *as* an instance of a class $C_1 = \{s \mid endangered\_pedestrian(s) = 1\}$ or *as* an instance of $C_2 = \{s \mid endangered\_pedestrian(s) = 0\}$. Note that C_2 is a rather broad category: it includes elephants, ottoman sofas, and heaps of salted pistachios, along with everything else in the universe (except, of course, some pedestrians). I would argue that the ability of such a pedestrian

avoidance system to "just see" is very limited, although it is not to be dismissed: it is not blind, merely egregiously single-minded.

In contrast, the ability of a human driver to "just see" is far more advanced than that of a pedestrian-avoidance module, because a human can interpret any given scene in a greater variety of ways: he or she can harbor a much larger number of concepts and can carry out more kinds of tasks. The human ability to "just see" is, however, very far from exhausting the range of conceivable possibilities. Think of a super-observer whose visual system is not encumbered by an attention bottleneck and who can perceive in a typical Manhattan scene (say) the location and disposition of every visible building and street fixture and can simultaneously track every unattached object, including chewing gum wrappers and popcorn kernels, as well as discern the species and the sex of every animal within sight, including pigeons, pedestrians, and the occasional rat.

A being with such powers of observation would be very good at "seeing as": for instance, should it have had sufficient experience in outer space travel, it may be capable of seeing the street scene *as* a reenactment of a series of collisions among rock and ice fragments in a particular cubic kilometer of the Oort cloud on January 1, 0800 hours UTC, 2008 CE, which it happened to have viewed while on a heliopause cruise. Equally importantly, however, it would also be very good at "just seeing" — a non-action[6] in which it can indulge merely by letting the seething mass of categorization processes that in any purposive visual system vie for the privilege of interpreting the input *be* the representation of the scene, without allowing any one of them to gain the upper hand.[7]

Note that although the evolution of visual systems may well be driven by their role in supporting action and by their being embodied in active, purposive agents (Noë, 2004), once the system is in place no action is required for it to "just see" (Edelman, 2006). When not driven by the demands of a specific task, the super-observer system just imagined may see its surroundings *as* nothing in particular, yet its visual experience would be vastly richer than ours, because of the greater number of aspects made explicit in (and therefore potential distinctions afforded by) its representation of the scene.

This brings us to a key realization: rather than conceptual, purposive, and interpretation-driven, visual experience, whether rich or impoverished, is representational. As Wittgenstein (1958) noted, "To interpret is to think, to do something; seeing is a state."[8] We are now in a position to elaborate on this observation: seeing is a *representational* state

(Edelman, 2002; for a detailed discussion, see Edelman, 2008, sec. 5.7 and 9.4).

1.3 A closer look at "seeing as"

The foregoing discussion suggests that to understand the computational nature and possible range of pure visual experience, or "just seeing," we must first understand the nature of conceptual vision, or "seeing as," of which "just seeing" is a kind of by-product (at least in evolved rather than engineered visual systems). In the early years of principled computational study of vision, the efforts to understand "seeing as" focused on charting the possible paths leading from raw image data to seeing the world as a spatial arrangement of surfaces, volumes, and, eventually, objects (Aloimonos and Shulman, 1989; Marr, 1982; Marr and Nishihara, 1978). The key observation, due to Marr, was that this goal could be approached by processing the input so as to *make explicit* (Marr, 1982, pp.19-24) the geometric structure of the environment that is implicitly present in the data. This research program thus amounts to an attempt to elucidate how the geometry of the world could be reconstructed from the visual input.

Both the feasibility of and the need for an explicit and sweeping reconstruction of the geometry of the visual world have been subsequently questioned (Aloimonos et al., 1988; Bajcsy, 1988). Noting that biological vision is purposive and active, researchers proposed that computer vision too should aim at serving certain well-defined goals such as navigation or recognition rather than at constructing a general-purpose representation of the world. Moreover, a visual system should actively seek information that can be used to further its goals. This view rapidly took over the computer vision community. At present, all applied work in computer vision is carried out within the purposive framework; the role of active vision is especially prominent in robotics.

From the computational standpoint, this development amounted to shifting the focus of research from "inverse optics" approaches (Bertero et al., 1988), which aim to recover the solid geometry of the viewed scene, to managing feature-based evidence for task-specific hypotheses about the input (Edelman and Poggio, 1989). This shift occurred in parallel with the gradual realization that the prime candidate framework for managing uncertainty — graphical models, or Bayes networks — is ubiquitous in biological vision (Kersten et al., 2004; Kersten and Yuille, 2003; Knill and Richards, 1996), as it is, indeed, in cognition in general

(Chater et al., 2006). Importantly, the Bayesian framework allows for a seamless integration of bottom-up data with prior assumptions and top-down expectations, without which visual data are too underdetermined to support reliable decision-making (Marr, 1982; Mumford, 1996). Such integration is at the core of the most promising current approaches to object and scene vision (Fei-Fei et al., 2003; Freeman, 1993; Torralba et al., 2003), including the explicitly generative "analysis by synthesis" methods (Yuille and Kersten, 2006).

1.4 The problems with "seeing as"

Both major approaches to vision — scene reconstruction and purposive processing — run into problems when taken to the limit. On the one hand, vision considered as reconstruction is problematic because complete recovery of detailed scene geometry is infeasible, and because a replica of the scene, even if it were available, would not in fact further the goal of conceptual interpretation — seeing the scene as something (Edelman, 1999). On the other hand, extreme purposive vision is problematic because a system capable of performing seventeen specific tasks may still prove to be effectively blind when confronted with a new, eighteenth task (Intrator and Edelman, 1996). To better appreciate the issues at hand, let us consider three factors in the design of a visual system: the role of the task, the role of the context in which a stimulus appears, and the role of the conceptual framework within which vision has to operate.

1.4.1 The role of the task

Given that biological visual systems are selected (and artificial ones engineered) not for contemplation of the visible world but for performance in specific tasks, it would appear that the purposive approach is the most reasonable one to pursue — provided that the list of visual tasks that can possibly matter to a given system is manageably short. Deciding whether the purposive approach is feasible as a general strategy for vision reduces therefore to answering the question "What is vision for?" In practice, however, the need to develop a taxonomy of visual tasks has not been widely recognized in vision research (the works of Marr (1982), Ballard (1991), Aloimonos (1990), and Sloman (1987, 1989, 2006) are some of the rare exceptions).

The unavailability of a thorough, let alone complete, taxonomy of visual tasks has a reason other than the sheer tediousness of taxonomic work. The reason is this: insofar as vision is to be useful to an active agent (biological or engineered) in confronting the real world, it must be *open-ended*. Specifying ahead of time the range of tasks that a visual system may need to face is impossible because of a very general property of the universe: the open-endedness of the processes that generate complexity — especially the kind of complexity that pervades the biosphere (Clayton and Kauffman, 2006).

The relentless drive toward higher complexity in ecosystems can be illustrated by the simple example of a situation in which a predator must decide between two species of prey: inedible (toxic) "models," and edible "mimics" (Tsoularis, 2007). The resort to mimicry by the edible prey presents a computational challenge to the predator, whose perceptual system must learn to distinguish among increasingly similar patterns presented by the prey, on the pain of indigestion, starvation, and possibly death.[9] The mimic species faces a similar perceptual challenge (albeit dissimilar consequences of a wrong decision) in mate choice.[10] Crucially for the evolution of a visual system that is thrown into the midst of such a computational arms race, mimicry situations typically involve "rampant and apparently easy diversification of mimetic patterns" (Joron, 2003).

Note that counting new perceptual distinctions as new "tasks" in the preceding example falls squarely within the computational complexity framework based on VC dimension, which is all about counting ways to classify the data into distinct categories. Complexity theory is neutral with respect to the actual methods whereby classification can be learned and progressively finer perceptual distinctions supported. Of the many such methods (Hastie et al., 2001), I mention here one of the simplest, the Chorus of Prototypes (Edelman, 1998, 1999). According to this method, the representation space into which new stimuli are cast and in which the categorization decision is subsequently made is spanned by the outputs of filter-like units tuned to some of the previously encountered stimuli (the "prototypes") — a representation that can be learned simply by "imprinting" units newly recruited one after another with select incoming filter patterns.

Employing the terminology introduced earlier, we may observe that a stimulus presented to such a system is thereby simultaneously "seen as" each of the existing prototypes (in a graded rather than all-or-none sense, because the responses of the prototype units are graded). The denser the coverage of a given region of the stimulus space by prototypes,

the finer the discrimination power that is afforded in that region to the system by the vector of similarities to the prototypes (and the higher the VC dimension of the system). Crucially, if discrimination is deferred, the mere representation of the stimulus by the outputs of the prototype-tuned filters still amounts to "just seeing" it — that is, to having a visual experience whose richness is determined by the dimensionality and, very importantly, by the spatial structure and the prototype composition of the representation space.

Why do the structure and the composition of the representation space spanned by the system's conceptual back end matter so much? Although any input scene is necessarily also represented in the front end (as a vector of pixel values or photoreceptor activities), this more primitive representation does not make explicit various behaviorally and conceptually consequential aspects of the scene. The human visual system harbors both raw (pixel-like) representations and a great variety of structured ones, while a pedestrian detection system may only need the former; this is why a human is much better not only at seeing the visual world *as* a profusion of objects, but also at "just seeing" it (insofar as he or she can make sure that "seeing as" does not get in the way).[11]

1.4.2 The role of context

The runaway proliferation of visual tasks, which as noted above include the distinctions that need to be made among various stimuli, stems not only from the complexity of the stimuli by themselves, but also from the diversity of the contexts in which they normally appear. This latter, contextual complexity figures prominently in what Sloman (1983, p.390) called "the horrors of the real world" that beset computer vision systems.

One problem posed by real-world scenes is that recognizable objects, if any, tend to appear in the wild against highly cluttered backgrounds (Oliva and Torralba, 2007). I illustrate this point with two photographs: Figure 1.1, top, shows an urban scene in which some common objects (a car, a cat, a house) appear at a medium distance; Figure 1.1, bottom, shows a close-up of a rain-forest floor centered on a snail clinging to a rotting mango. Reliable detection (let alone recognition) of objects in such scenes was impossible until recently in computer vision. Highly purposive systems limited to dealing with a small number of object classes are now capable of finding their target objects in cluttered scenes, by employing Bayesian methods that combine bottom-up and top-down cues (Torralba et al., 2003; Weber et al., 2000; Yuille and Kersten, 2006).

Fig. 1.1. Two real-world scenes. *Top:* an urban environment, mid-distance. *Bottom:* a natural environment, close-up.

Being class-specific, these methods cannot, however, solve the wider problem posed by real-world clutter: the impossibility of constructing an exhaustive and precise description of any scene that is even halfway interesting. The best that a targeted recognition system can hope for is attaining a sparse, conceptual description, as when the arid pasture scene of Figure 1.2, top, is mapped into the set of spatially anchored labels shown at the bottom. By now, computer vision researchers seem to have realized that reconstructing the detailed geometry of such scenes, in which the shape and pose of every pebble and the disposition of every blade of grass is made explicit (as in the $2\frac{1}{2}D$ sketch of Marr (1982) or the intrinsic images of Barrow and Tenenbaum (1978)), is not feasible (Barrow and Tenenbaum, 1993; Dickinson et al., 1997).

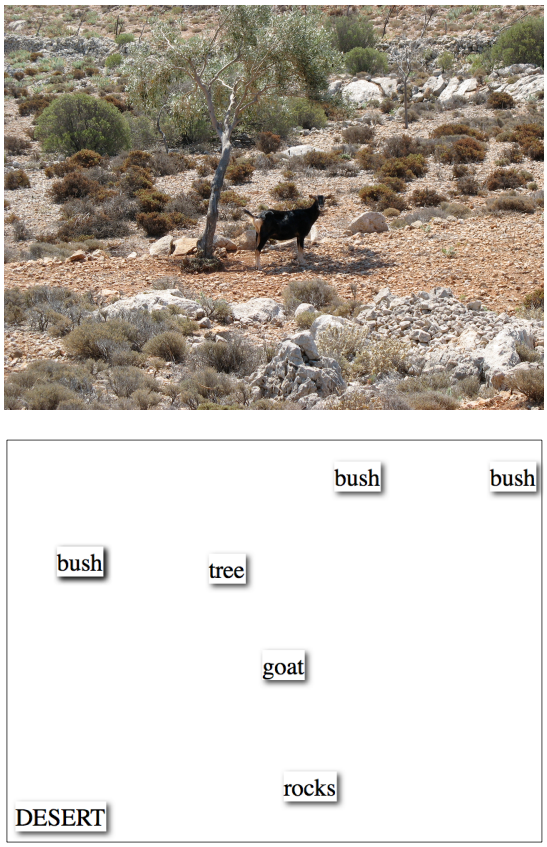

Fig. 1.2. Two versions of a real-world scene. *Top:* a natural environment. *Bottom:* the same natural scene, represented by spatially anchored conceptual labels.

Our visual experience would be impoverished indeed if we were capable of seeing the scenes of Figures 1.1 and 1.2 only "as" parked car, rotting mango, or grazing goat, respectively.[12] These photographs[13] strike us as replete with visual details. Most of these details are, however, "just seen," not "seen as" anything; computer vision systems too need not attempt the humanly impossible when confronted with real-world scenes. Matching the complexity of a human experience of the visual world is a realistic goal, and is challenging enough. As we saw earlier, representations that would make such a match possible are also likely to support highly sophisticated purposive vision.

1.4.3 The role of conceptual knowledge

As just noted, purposive visual systems can only deliver scene descriptions that are (1) sparse, and (2) conceptual. The second of these properties, or rather limitations, is no less important than the first one (which I discussed briefly above). Restricting the representations derived from scenes to being conceptual amounts to imposing a severe handicap on the visual system. At the level of description with which human "just seeing" resonates, the natural visual world is *ineffable*, in that a vast majority of its "aspects" are not statable in a concise linguistic form; indeed, most are non-conceptual (Clark, 2000, p.162).[14] Correspondingly, philosophers point out that "Perceptual experience has a richness, texture and fineness of grain that [conceptual] beliefs do not and cannot have" (Bermúdez, 1995; see also Akins, 1996; Villela-Petit, 1999).

When a set of conceptual labels is applied to a visual scene and is allowed to take over the representation of that scene, the ineffability issue gives rise to two sorts of problems. The first problem stems from the poverty of conceptual labels; earlier in this section I used Figure 1.2 to illustrate the extent to which a conceptual interpretation of a scene is impoverished relative to its image. The second problem arises when one tries to decide where exactly to place the boundary between areas corresponding to each two adjacent labels — precisely the task with which users of interactive scene labeling applications such as LabelMe (Russell et al., 2007) are charged.

The common mistake behind various attempts to develop the ultimate algorithm for scene segmentation, whether using image data or input from a human observer, is the assumption that there is a "matter of fact" behind segmentation.[15] For natural scenes, segmentation is in the eye of the beholder: the same patch may receive different labels from different users or from the same user engaged in different tasks (cf. Figure 1.3), or no label at all if it is too nondescript or if it looks like nothing familiar.[16] To a visually sophisticated observer, a complex natural scene would normally appear as continuous canvas of rich experience, rather than as a solved puzzle with labeled pieces. Even if nothing in the scene is "seen as" something familiar, the whole, and whatever fleeting patterns that may be discerned in it, can always be "just seen" in the sense proposed above.

To summarize, the major challenges that arise in the design of an advanced visual system — adapting to diverse tasks, dealing with realistic contexts, and preventing vision from being driven exclusively by con-

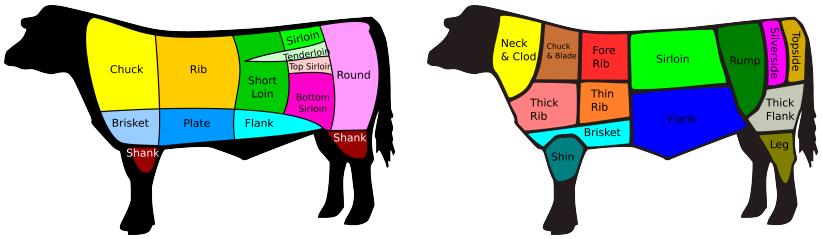

Fig. 1.3. Concepts that may affect scene segmentation are not necessarily universal, as illustrated metaphorically by these butchers' diagrams, which compare the US cuts of beef (left) to the British cuts (right). Ask an English butcher for a piece of beef tenderloin, and you will not be understood.

ceptual knowledge — can all be met in the same way. This middle way, which calls for fostering properly structured intermediate representations while avoiding the symmetrical excesses of full geometric reconstruction and full conceptual interpretation, corresponds precisely to "just seeing." Somewhat paradoxically, therefore, it is "just seeing" that saves the day for "seeing as."

1.5 Some parallels with biological vision

In computer vision, the discussion of what it means to see can afford to be normative, in suggesting what a good visual system should be doing. In biological vision, in contrast, the first order of business is finding out what it is that living visual systems actually do. What a visual system does depends on the animal in which it is embodied and on the ecological niche in which the animal resides. For instance, in the behavioral repertoire of the bay scallop, escaping danger by rapidly pulling the shell shut occupies a prominent place. The scallop's visual system, which is fed information from the many tiny eyes that line the rim of its mantle, triggers the escape reflex in response to the onset of a shadow (Hartline, 1938; Wilkens and Ache, 1977).

Even when the shadow is in fact cast by a cuttlefish on the prowl, it would be unparsimonious to assume that the scallop sees it *as* a manifestation of the concept cuttlefish: scallops are simply wired to propel themselves away from shadows (just as frogs are preset to snap at dark moving dots that may or may not be flies, and flies are compelled to chase other dark moving dots).[17] Near the other end of the spectrum of visual sophistication, the primate visual system (Kremers, 2005) incor-

porates, in addition to a multitude of reflexes, a variety of classification- and action-related functions.

The now familiar contrast between "just seeing" and "seeing as" can be interpreted in terms of a major distinction that exists among the various functions of the primate visual system. In anatomical terms, it corresponds to the distinction between mesencephalic (midbrain) and telencephalic (forebrain) visual systems. A key part of the former is the superior colliculus (King, 2004): a structure in the midbrain's "roof" or tectum, where sensory (visual, auditory, and somatic), motor, and motivational representations are brought together in the form of spatially registered maps (Doubell et al., 2003).

With only a slight oversimplification, it may be said that the superior colliculus (SC) is the engine of purposive vision: if the animal is motivated to reach out to a stimulus that its eyes fixate, the action is coordinated by SC neurons (Stuphorn et al., 2000). It is the sparing of subcortical structures including the thalamus and the SC that supports blindsight (Stoerig and Cowey, 1997) and makes possible the persistence of a primitive kind of visual consciousness (Merker, 2007) in patients with severe cortical damage.

The association networks of concepts (visual and other) that make primate cognition so powerful are distilled from long-term memory traces of the animal's experiences. Because these networks reside in the forebrain (Merker, 2004), mesencephalic vision, which bypasses the isocortical structures in primates, is non-conceptual, although the purposive behavior that it can support may be quite flexible (insofar as its planning involves integrating information from multiple sources, including context and goals). As such, the midbrain visual system is not good at "just seeing" — a function that, as I argued earlier, is built on top of the capacity for "seeing as."

In primates, the capacity for "seeing as" is supported by isocortical structures that consist of the primary visual areas in the occipital lobe and the high-level areas in the temporal and parietal lobes (Rolls and Deco, 2001), and the frontal lobe, the visual functions of which include exerting contextual influence on the interpretation of the viewed scene (Bar, 2004) and active vision or foresight (Bar, 2007). In computational terms, the cortical visual system represents the scene by the joint firing of banks of neurons with graded, overlapping receptive fields, which are coarsely tuned to various "objects" (which may be conceptually quite sophisticated) and are modulated by top-down signals (Edelman, 1999). By virtue of having a cortical visual system — over and above (literally)

the vertebrate standard-issue one in the midbrain — primates can see the world *as* so many different things, as well as just see it.

1.6 Conclusions

> We find certain things about seeing puzzling, because we do not find the whole business of seeing puzzling enough.[18]
>
> Ludwig Wittgenstein (1889-1951)

Contrary to the widespread but tacit assumption in the sciences of vision, having a well-developed sense of sight corresponds to more than the ability to recognize and manipulate objects and to interpret and navigate scenes. The behavioral, neurobiological, and computational insights into the workings of primate vision that emerged in the past two decades go a long way towards characterizing the component that has hitherto been missing from most accounts of vision. The missing component is the capacity for having rich visual experiences.

In a concrete computational sense, visual experience is not merely an epiphenomenon of visual function. A profound capacity for perceptual contemplation goes together with the capacity for seeking out flexible, open-ended mappings from perceptual stimuli to concepts and to actions. In other words, the ability to see the world *as* an intricate, shifting panoply of objects and affordances — an oft-discussed mark of cognitive sophistication (Hofstadter, 1995) — is coextensive with the ability to "just see."

From a computational standpoint, this ability requires that the visual system maintain versatile intermediate representations that (1) make explicit as wide as possible a variety of scene characteristics, and (2) can be linked in a flexible manner to a conceptual system that is capable of growing with need and experience. These requirements transcend the traditional goals of high-level vision, which are taken to be the ability to recognize objects from a fixed library and to guess the gist of scenes. The visual world is always more complex than can be expressed in terms of a fixed set of concepts, most of which, moreover, only ever exist in the imagination of the beholder.

Luckily, however, visual systems need not explain the world — they only need to resonate to it in various useful ways (Gibson, 1979; Sloman, 1989). Anticipating the idea of O'Regan (1992) and O'Regan and Noë (2001) who argued that the world is its own best representation, Reitman

et al. (1978, p.72) observed that "The primary function of perception is to keep our internal framework in good registration with that vast external memory, the external environment itself." To be able to resonate with the virtually infinite perceivable variety of what's out there — quoting William Blake, "to see a world in a grain of sand" — an advanced visual system should therefore strive for the richness of the measurement front end, the open-endedness of the conceptual back end,[19] and the possibility of deferring conceptualization and interpretation in favor of just looking.[20]

Acknowledgments

Thanks to Melanie Mitchell for inviting me to a Santa Fe Institute workshop ("High-Level Perception and Low-Level Vision: Bridging the Semantic Gap," organized by M. Mitchell and G. Kenyon) that prompted me to rethink answers to questions in the computational neurophenomenology of vision that preoccupied me for some time. Thanks also to Tony Bell and to David Ackley for their remarks following my talk at SFI, and to Melanie, Tomer Fekete, and Catalina Iricinschi for commenting on a draft of this chapter.

Notes

1 *Philosophical Investigations*, (Wittgenstein, 1958, II,xi).

2 Although intuition is never to be trusted blindly, we must use it as a starting point in a process of formalization, because the notion of seeing is itself inherently intuitive rather than formal to begin with. In that, it is similar to the notion of effective computation, which is invoked by the Church-Turing Thesis.

3 For a discussion of the nominal dimensionality of continuous measurement spaces and the actual dimensionality of data sets mapped into such spaces, see Edelman (1999). The same topics are treated in terms of persistent homology theory by Fekete et al., *Arousal increases the representational capacity of cortical tissue* (2008, submitted).

4 A set $\mathcal{S}$ is shattered by the binary concept class $\mathcal{C}$ if for each of the $2^{|\mathcal{S}|}$ subsets $s \subseteq \mathcal{S}$ there is a concept $f \in \mathcal{C}$ that maps all of s to 1 and $\mathcal{S} - s$ to 0. The analytical machinery of VC dimension can be extended to deal with real-valued concepts: for a class of real-valued function $g : \mathcal{S} \to \mathbb{R}$, the VC dimension is defined to be that of the indicator class $\{I(g(s) - \beta > 0)\}$ where β takes values over the range of g (Hastie et al., 2001). An extension to multiple-valued concepts is also possible (Bradshaw, 1997).

5 A fanciful literary example of a cognitive system crippled by its own enormous capacity for individualizing concepts can be found in the short story *Funes the Memorious* by Jorge Luis Borges (1962); a real case has

been described by A. Luria in *The Mind of a Mnemonist* (Harvard: 1968).

6 For the concept of non-action, or *wu wei*, see Loy (1985).

7 Because the activation levels of conceptual representations are graded, there exists a continuum between "just seeing" and "seeing as" (I am grateful to Melanie Mitchell for pointing out to me this consequence of the approach to vision outlined in this paper). A distributed conceptual system (e.g., the Chorus of Prototypes model of visual recognition and categorization; Edelman, 1999) may position itself along this continuum by controlling its dynamics — in the simplest case, a single "temperature" parameter (Hofstadter and Mitchell, 1995).

8 Wittgenstein's observation concerning the nature of vision may have been anticipated by Aristotle in *Metaphysics* (350 B.C.E., IX,8): "In sight the ultimate thing is seeing, and no other product besides this results from sight."

9 Famous last words of a mistaken predator: "Oops, it sure *looked* tasty."

10 Famous last words of a too undiscriminating sex partner seeker: "Care for a dance, mate?", spoken to a trigger-happy alien that *looked* like a member of one's opposite sex.

11 The distinction between the kinds of experience afforded by low-level, pixel-like representations and high-level ones spanned by similarities to prototypes is crucial for understanding how the so-called "hard problem" of consciousness (Chalmers, 1995), which pertains to visual qualia, is fully resolved by Smart (2004): "Certainly walking in a forest, seeing the blue of the sky, the green of the trees, the red of the track, one may find it hard to believe that our qualia are merely points in a multidimensional similarity space. But perhaps that is what *it is like* (to use a phrase that can be distrusted) to be aware of a point in a multidimensional similarity space." Briefly, qualia that exist as points in a *structured* space (such as the one spanned by a set of prototype-tuned units; Edelman, 1999) can pertain to any and all aspects of the stimulus (over and above mere local intensities represented at the "pixel" level). Smart's insight thus accounts in a straightforward computational manner for the supposedly mysterious nature of perceptual experience.

12 The approach to scene "description" illustrated in Figure 1.2 has been lampooned by René Magritte in paintings such as *From One Day to Another* and *The Use of Speech* (Edelman, 2002).

13 High-resolution originals of the photographs in Figures 1.1 and 1.2 are available from the author by request.

14 To the extent that non-human animals and prelinguistic infants are capable of conceptual cognition (Smith and Jones, 1993; Vauclair, 2002), concepts need not be linguistic. If and when available, language does, of course, markedly boost the ability to think conceptually (Clark, 1998; Dennett, 1993).

15 The Platonist notion that there exists an absolute truth about the conceptual structure of world "out there" that only needs to be discovered is not peculiar to theories of vision: it has been the mainstay of theoretical linguistics for decades. This notion underlies the distinction made by Householder (1952) between what he termed "God's truth" and "hocus-pocus" approaches to theorizing about the structure of sentences, the former one being presumably the correct choice. Although it still

survives among the adherents of Chomsky's school of formal linguistics, the idea that every utterance possesses a "God's truth" analysis seems to be on its way out (Edelman and Waterfall, 2007).

16 The few exceptions to this general pattern are provided by scenes in which a prominent object is foregrounded by a conjunction of several cues, as when a horse is seen galloping in a grassy field; such images figure prominently in computer vision work on scene segmentation, e.g., that of Borenstein and Ullman (2002).

17 In contrast to scallops, which can act on what they see but not classify it in any interesting sense, the HabCam computer vision system built by Woods Hole marine biologists, which carries out a high-resolution scan of the ocean floor (Howland et al., 2006), can classify and count scallops in the scenes that it registers. This undoubtedly qualifies it as capable of seeing scallops *as* such.

18 *Philosophical Investigations*, (Wittgenstein, 1958, II,xi).

19 An intriguing computational mechanism that seems capable of implementing an open-ended representational system is the liquid-state machine of Maass et al. (2003) (for a recent review, see Maass, 2007). The power of LSMs to support classification is related to that of support-vector machines (Cortes and Vapnik, 1995).

20 With regard to the virtues of "just looking," consider the following piece of inadvertent propaganda for *wu wei*: "Don't just do something, stand there!" — White Rabbit to Alice in the film *Alice in Wonderland* (1951).

References

Akins, K. (1996). Of sensory systems and the 'aboutness' of mental states. *Journal of Philosophy*, XCIII:337–372.

Aloimonos, J. Y. (1990). Purposive and qualitative vision. In *Proc. AAAI-90 Workshop on Qualitative Vision*, pages 1–5, San Mateo, CA. Morgan Kaufmann.

Aloimonos, J. Y. and Shulman, D. (1989). *Integration of visual modules: an extension of the Marr paradigm*. Academic Press, Boston.

Aloimonos, J. Y., Weiss, I., and Bandopadhay, A. (1988). Active vision. *Intl. J. Computer Vision*, 2:333–356.

Aristotle (350 B.C.E.). Metaphysics. Available online at http://classics.mit.edu/Aristotle/metaphysics.html.

Bajcsy, R. (1988). Active perception. *Proc. IEEE*, 76(8):996–1005. special issue on Computer Vision.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48:57–86.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5:617–629.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11:280–289.

Barrow, H. G. and Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In Hanson, A. R. and Riseman, E. M., editors, *Computer Vision Systems*, pages 3–26. Academic Press, New York, NY.

Barrow, H. G. and Tenenbaum, J. M. (1993). Retrospective on "Interpreting line drawings as three-dimensional surfaces". *Artificial Intelligence*, 59:71–80.

Basri, R. and Ullman, S. (1988). The alignment of objects with smooth surfaces. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 482–488, Tarpon Springs, FL. IEEE, Washington, DC.

Baum, E. B. and Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, 1:151–160.

Bermúdez, J. L. (1995). Non-conceptual content: From perceptual experience to subpersonal computational states. *Mind and Language*, 10:333–369.

Bertero, M., Poggio, T., and Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1986). Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. In *18th annual ACM symposium on theory of computing*, pages 273–282.

Borenstein, E. and Ullman, S. (2002). Class specific top down-segmentation. In Heyden, A., editor, *Proceedings of the European Conference on Computer Vision*, volume 2351 of *Lecture Notes in Computer Science*, pages 110–122.

Borges, J. L. (1962). *Ficciones*. Grove Press, New York. Translated by A. Bonner in collaboration with the author.

Bradshaw, N. P. (1997). The effective VC dimension of the n-tuple classifier. In *Proc. Artificial Neural Networks – ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 511–516, Berlin. Springer.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2:200–219.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10:287–291.

Clark, A. (1998). Magic words: How language augments human computation. In Carruthers, P. and Boucher, J., editors, *Language and thought: Interdisciplinary themes*, pages 162–183. Cambridge University Press, Cambridge.

Clark, A. (2000). *A theory of sentience*. Oxford University Press, Oxford.

Clayton, P. and Kauffman, S. A. (2006). Agency, emergence, and organization. *Biology and Philosophy*, 21:501–521.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Dennett, D. C. (1993). Learning and labeling. *Mind and Language*, 8:540–547.

Dickinson, S., Bergevin, R., Biederman, I., Eklundh, J., Munck-Fairwood, R., Jain, A., and Pentland, A. (1997). Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15:277–292.

Doubell, T. P., Skaliora, T., Baron, J., and King, A. J. (2003). Functional connectivity between the superficial and deeper layers of the superior colliculus: an anatomical substrate for sensorimotor integration. *Journal of Neuroscience*, 23:6596–6607.

Edelman, S. (1993). On learning to recognize 3D objects from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:833–837.

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498.

Edelman, S. (1999). *Representation and recognition in vision*. MIT Press, Cambridge, MA.

Edelman, S. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences*, 6:125–131.

Edelman, S. (2006). Mostly harmless: review of *Action in Perception* by Alva Noë. *Artificial Life*, 12:183–186.

Edelman, S. (2008). *Computing the mind: how the mind really works*. Oxford University Press, New York.

Edelman, S. and Intrator, N. (2002). Models of perceptual learning. In Fahle, M. and Poggio, T., editors, *Perceptual learning*, pages 337–353. MIT Press.

Edelman, S. and Poggio, T. (1989). Representations in high-level vision: reassessing the inverse optics paradigm. In *Proc. DARPA Image Understanding Workshop*, pages 944–949, San Mateo, CA. Morgan Kaufman.

Edelman, S., Ullman, S., and Flash, T. (1990). Reading cursive handwriting by alignment of letter prototypes. *Intl. J. Computer Vision*, 5:303–331.

Edelman, S. and Waterfall, H. R. (2007). Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews*, 4:253–277.

Fei-Fei, L., Fergus, R., and Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV-2003*.

Freeman, W. T. (1993). Exploiting the generic view assumption to estimate scene parameters. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 347–356, Washington, DC. IEEE.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

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

Hartline, H. K. (1938). The discharge of impulses in the optic nerve of *Pecten* in response to illumination of the eye. *J. Cell. Comp. Physiol.*, 2:465=478.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Hofstadter, D. R. (1995). On seeing A's and seeing As. *Stanford Humanities Review*, 4:109–121.

Hofstadter, D. R. and Mitchell, M. (1995). The Copycat project: a model of mental fluidity and analogy-making. In Hofstadter, D. R., editor, *Fluid Concepts and Creative Analogies*, chapter 5, pages 205–265. Basic Books, NY.

Householder, F. W. (1952). Review of Harris, Zellig S., Methods in Structural Linguistics. *International Journal of American Linguistics*, 18:260–268.

Howland, J., Gallager, S., Singh, H., Girard, A., Abrams, L., Griner, C., Taylor, R., and Vine, N. (2006). Development of a towed survey system for deployment by the fishing industry. *Oceans*, pages 1–5.

Huttenlocher, D. P. and Ullman, S. (1987). Object recognition using alignment. In *Proceedings of the 1st International Conference on Computer Vision*, pages 102–111, London, England. IEEE, Washington, DC.

Intrator, N. and Edelman, S. (1996). How to make a low-dimensional representation suitable for diverse tasks. *Connection Science*, 8:205–224.

Joron, M. (2003). Mimicry. In Cardé, R. T. and Resh, V. H., editors, *Encyclopedia of Insects*, pages 714–726. Academic Press, New York.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304.

Kersten, D. and Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13:1–9.

King, A. J. (2004). The superior colliculus. *Current Biology*, 14:R335–R338. A primer.

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

Kremers, J., editor (2005). *The Primate Visual System*. John Wiley & Sons, New York.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395.

Loy, D. (1985). Wei-wu-wei: Nondual action. *Philosophy East and West*, 35:73–87.

Maass, W. (2007). Liquid computing. In *Proceedings of the CiE'07 Conference: Computability in Europe 2007*, Lecture Notes in Computer Science, Berlin. Springer.

Maass, W., Natschläger, T., and Markram, H. (2003). Computational models for generic cortical microcircuits. In Feng, J., editor, *Computational Neuroscience: A Comprehensive Approach*, chapter 18, pages 575–605. CRC-Press, Boca Raton, FL.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, 200:269–294.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488.

Merker, B. (2004). Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex*, 40:559–576.

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30:63–81.

Mumford, D. (1996). Pattern theory: a unifying perspective. In Knill, D. and Richards, W., editors, *Perception as Bayesian Inference*. Cambridge Univ. Press, Cambridge, UK.

Noë, A. (2004). *Action in Perception*. MIT Press, Cambridge, MA.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11:520–527.

O'Regan, J. K. (1992). Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:883–917.

Reitman, W., Nado, R., and Wilcox, B. (1978). Machine perception: what makes it so hard for computers to see? In Savage, C. W., editor, *Perception and cognition: issues in the foundations of psychology*, volume IX of *Minnesota studies in the philosophy of science*, pages 65–87. University of Minnesota Press, Minneapolis, MN.

Rolls, E. and Deco, G. (2001). *Computational Neuroscience of Vision*. Oxford University Press, New York.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. T. (2007). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*. DOI: 10.1007/s11263-007-0090-8.

Sloman, A. (1983). Image interpretation: The way ahead? In Braddick, O. J. and Sleigh, A. C., editors, *Physical and Biological Processing of Images*, Springer Series in Information Sciences, pages 380–401. Springer Verlag, Berlin Heidelberg New York.

Sloman, A. (1987). What are the purposes of vision? CSRP 066, University of Sussex.

Sloman, A. (1989). On designing a visual system (towards a Gibsonian computational model of vision). *J. of Experimental and Theoretical Artificial Intelligence*, 1:289–337.

Sloman, A. (2006). Aiming for more realistic vision systems? COSY-TR 0603, University of Birmingham, School of Computer Science.

Smart, J. J. C. (2004). The identity theory of mind. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. Stanford University. Available online at http://plato.stanford.edu/archives/fall2004/entries/mind-identity/.

Smith, L. B. and Jones, S. (1993). Cognition without concepts. *Cognitive Development*, 8:181–188.

Stoerig, P. and Cowey, A. (1997). Blindsight in man and monkey. *Brain*, 120:535–559.

Stuphorn, V., Bauswein, E., and Hoffmann, K. P. (2000). Neurons in the primate superior colliculus coding for arm movements in gaze-related coordinates. *Journal of Neurophysiology*, 83:1283–1299.

Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proc. IEEE Intl. Conference on Computer Vision (ICCV)*, pages 273–281, Nice, France.

Tsoularis, A. (2007). A learning strategy for predator preying on edible and inedible prey. *Acta Biotheoretica*, 55:283–295.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag, Berlin.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.

Vauclair, J. (2002). Categorization and conceptual behavior in nonhuman primates. In Bekoff, M., Allen, C., and Burghardt, G., editors, *The Cognitive animal*, pages 239–245. MIT Press, Cambridge, MA.

Villela-Petit, M. (1999). Cognitive psychology and the transcendental theory of knowledge. In Petitot, J., Varela, F. J., Pachoud, B., and Roy, J.-M., editors, *Naturalizing phenomenology: issues in contemporary phenomenology and cognitive science*, pages 508–524. Stanford University Press, Stanford, CA.

Weber, M., Welling, M., and Perona, P. (2000). Unsupervised learning of models for recognition. In Vernon, D., editor, *Proceedings of the European Conference on Computer Vision*, volume 1842 of *Lecture Notes in Computer Science*, pages 18–32, Berlin. Springer.

Wilkens, L. A. and Ache, B. W. (1977). Visual responses in the central nervous system of the scallop *Pecten ziczac*. *Cellular and Molecular Life Sciences*, 33:1338–1340.

Wittgenstein, L. (1958). *Philosophical Investigations*. Prentice Hall, Englewood Cliffs, NJ, 3rd edition. Translated by G. E. M. Anscombe.

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10:301–308.