Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Neurobiology

Ambiguity and invariance: two fundamental challenges for visual processing

Nicole C Rust and Alan A Stocker

The visual system is tasked with extracting stimulus content (e.g. the identity of an object) from the spatiotemporal light pattern falling on the retina. However, visual information can be ambiguous with regard to content (e.g. an object when viewed from far away), requiring the system to also consider contextual information. Additionally, visual information originating from the same content can differ (e.g. the same object viewed from different angles), requiring the system to extract content invariant to these differences. In this review, we explore these challenges from experimental and theoretical perspectives, and motivate the need to incorporate solutions for both ambiguity and invariance into hierarchical models of visual processing.

Address
Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Room 302C, Philadelphia, PA 19104, USA

Corresponding author: Rust, Nicole C (nrust@sas.upenn.edu)

Current Opinion in Neurobiology 2010, **20**:382–388

This review comes from a themed issue on
Sensory systems
Edited by Kevan Martin and Kristin Scott

0959-4388/$ – see front matter

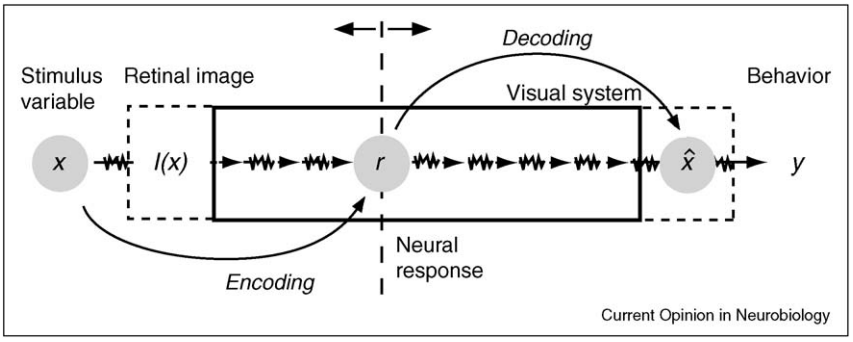
DOI 10.1016/j.conb.2010.04.013

Imagine you are in the park and about to meet your friend Suzie. As you look around, your visual system must identify Suzie within the rich environment that surrounds you. How does the visual system accomplish this feat? Visual processing is known to occur along an extended cascade of cortical processing stages [1] and thus the process of converting a stimulus into behavior is implemented across many different neural structures and brain areas (denoted by the small arrows; Figure 1). As a simplification, vision is often described as a two-stage encoding/decoding process. Encoding refers to the mapping of some stimulus variable, x (e.g. the identity of a person) onto the responses of one or more neurons, r. Conversely, decoding is the process that inverts the mapping, and thus arrives at an estimate $\hat{x}$ of the variable x given the neural response pattern r (e.g. inferring the identity of a person from the population response of neurons in a high-level visual area). Finally, the estimate $\hat{x}$ of the stimulus variable is used to direct a behavior y (e.g. calling out Suzie's name).

An elaboration of the encoding/decoding framework provides important insight into why visual processing is so challenging. At the initial sensory transduction stage, spatiotemporal patterns of light falling on the retina (the ''retinal image'', $I(x)$; Figure 1) are converted into neural signals by the photoreceptor array; the challenges of visual processing emerge from the fact that the value of nearly any stimulus variable x is only indirectly accessible via this representation. Specifically, identifying your friend Suzie requires your visual system to combine information across different points in the retinal image. Two computational challenges emerge from this basic requirement. The first, which we refer to as the ''ambiguity challenge'', results when different values of x produce identical or similar retinal images $I(x)$. For example, when viewed from a far distance, many individuals may produce a retinal image that is similar to the one produced by Suzie (Figure 2, left). To solve the ambiguity challenge, the visual system must make use of other sources of information. The second, which we refer to as the ''invariance challenge'', results when the same value of x produces different retinal images $I(x)$. For example, viewing Suzie in different positions and poses will produced very different $I(x)$ (Figure 2, right). To solve the invariance challenge, the visual system must associate the retinal images that contain the same value of x (e.g. Suzie) and differentiate these from the retinal images that contain different values of x (e.g. Layla or Lucy).

The problems of ambiguity and invariance are not unique to the task of identifying a person, but are encountered whenever the visual system attempts to estimate the value of a particular stimulus variable x from the environment. For example, when attempting to estimate motion direction, ambiguity can arise when two gratings, viewed through a small window, move in different directions but produce the same spatiotemporal light pattern (the ''aperture problem''), or when viewing conditions are noisy (e.g. driving in the fog). Estimation of motion direction also requires that direction be extracted in a manner invariant to the particular moving object or pattern. In other words, the natural and intrinsic goal of the visual system is not to provide a faithful account of the retinal image (like a camera) but rather to infer a (discrete or continuous) stimulus variable despite ambiguity and variation [2]. Below we review our current understanding of how the visual system deals with these two challenges.

Figure 1



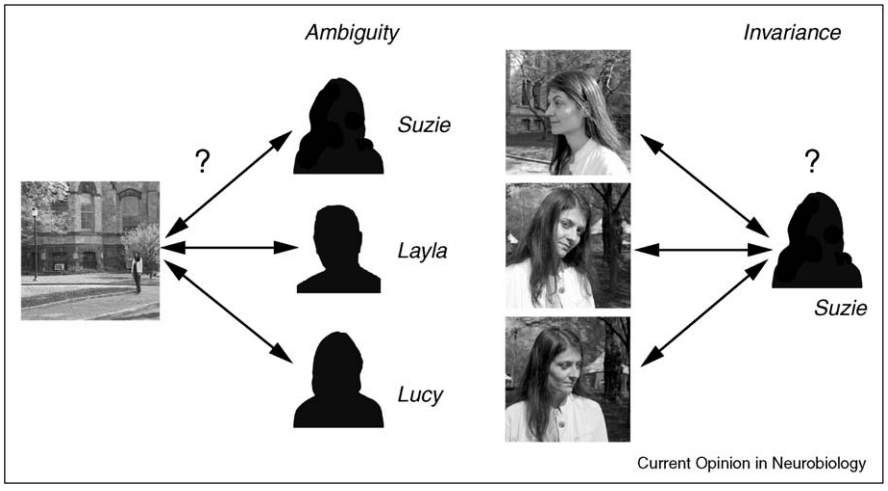
The encoding/decoding framework applied to visual processing. While the transformation of a stimulus variable x to a behavioral response y is implemented along a cascade of stages (denoted by the small arrows), visual processing is often simplified as a two-step encoding/decoding process. Encoding describes the mapping of x onto a neural response r whereas decoding describes the process of generating an estimate of the stimulus variable, $\hat{x}$, from r. Two fundamental challenges for vision, ambiguity and invariance, arise because x is only indirectly accessible via the spatiotemporal pattern of light intensity falling on the retina, $I(x)$ (the ''retinal image'').

The ''ambiguity challenge''

Perceptual scientists have long noted that the visual system considers sources of information beyond its visual input in order to unambiguously interpret the world. Helmholtz proposed that a percept is the visual system's ''best guess'' based on both prior knowledge and the visual information at hand [3]. For example, viewing a person in the park from far away may not provide sufficient visual information to unambiguously identify the person (Figure 2, left). However, knowing that you are about to meet Suzie in the park can lead to a reasonable guess of who that person could be [4•]. Note that we refer to ambiguity in a general sense, as uncertainty induced by a lack of information. This includes conditions where part of otherwise reliable information is missing (e.g. Suzie wearing sunglasses and a large hat that partially cover her face), as well as conditions where visual information is either degraded by noise or by a reduction in resolution.

Helmholtz's hypothesis can be naturally expressed in a probabilistic description within the encoding/decoding framework, which we refer to as the ''Bayesian observer model'' (see Box). A fundamental role of this model is to provide a quantitative description of how the visual system should combine noisy and ambiguous visual information with prior knowledge to produce an estimate of a stimulus variable (see [5] for a collection of early work). One example is a Bayesian observer model of speed estimation for moving patterns of different contrasts [6]. At low contrast, visual information is weak and

Figure 2



Examples of ambiguity and invariance. *Ambiguity*: Many different individuals, when viewed from afar, can generate a similar retinal image, resulting in visual information that is ambiguous with regard to the individual's identity. *Invariance*: The same individual can generate many different retinal images under different conditions and the visual system must associate retinal images with the same content invariant to these differences.

unreliable, and thus the uncertainty in determining speed based solely on the sensory information is high. Yet not all speeds occur equally often in the world; slow speeds are predominant [7]. A Bayesian observer model of speed estimation is a mathematical formulation of how sensory information (the likelihood function) and knowledge about the distribution of speeds in the word (the prior) should be optimally combined to determine the probability of all possible pattern speeds (the posterior). Selecting the most probable speed, this observer model predicts that moving patterns presented at low contrast should appear to move more slowly than they actually are moving, and this prediction is consistent with human perception [8]. The Bayesian observer model describes precisely how strongly prior knowledge should influence a percept at different levels of sensory uncertainty, and is a prediction of what a rational observer should do under these conditions.

Bayesian observer models have gained increasing support from the results of a variety of vision (or vision related) experiments including motion estimation [9], color perception [10], slant estimation [11], cue integration (e.g. [12–14]), visual search [15] and sensory-motor learning (e.g. [16]). Recent extensions to the Bayesian observer model have made it possible to reconstruct subjects' prior beliefs as well as their sensory uncertainty from behavioral data [9]. Such developments allow for important, quantitative tests, such as determining the extent to which subjects generalize their priors across different tasks.

The current popularity of the Bayesian observer model is due in part to the fact that it provides a simple and rational explanation for perceptual behavior under conditions of sensory uncertainty. The simplicity, however, comes with a price. In particular, two issues plague most of the Bayesian observer models proposed to date. First, encoding is often assumed to be a simplistic and abstract mapping between a stimulus variable x and population response r, to obtain a simple and tractable likelihood function (i.e. the generative model). One popular choice assumes that r is a direct mapping of x with additive Gaussian noise. Such simple encoding neglects the complexity of visual processing involved in mapping x to r through the retinal image $I(x)$ (described in more detail below). This leads us to a second issue: establishing a direct physiological implementation of a Bayesian observer model is difficult. A number of studies have proposed potential neural descriptions of how likelihood functions are explicitly (e.g. [17,18]) or implicitly (e.g. [19,20]) formed; how prior probabilities can be represented [21]; how likelihood functions can be multiplied (i.e. for cue integration) [19]; and how a percept $\hat{x}$ can be inferred [17]. Verifying these physiological models is likely to prove challenging [22]. Specifically, the Bayesian observer model is one instantiation of the abstract encoding/decoding framework introduced in the beginning of this review. As such, it is formulated for a specific encoding/decoding boundary, i.e. Bayesian inference is applied to a specific neural population r. However, such a boundary does not exist for the visual system, rather, visual processing takes place along a cascade of many processing stages. If the system as a whole performs Bayesian inference, it seems unlikely that any one stage in this cascade represents a single component of the Bayesian model (e.g. the prior) or performs one of the mathematical operations in isolation (e.g. multiplying the prior and the likelihood). Rather, one would expect that these operations and representations are distributed along the cascade, where neural and not mathematical constraints will determine their specific instantiations [23].

Box — Bayesian observer model

Neural noise and the uncertainty and ambiguity of the retinal image $I(x)$ with regard to the value of x motivate a probabilistic description of the encoding/decoding framework. Encoding can be characterized as the conditional probability distribution $p(r|x)$, describing the probability of observing a particular neural firing pattern r for a given value of the stimulus variable x. The response to each individual stimulus presentation is a sample of this probability distribution. Similarly, decoding can be described as the process that computes the probabilities for each value of x that this value has led to the observed firing pattern r, and then selects a value $\hat{x}$ appropriately. A decoder can directly compute these probabilities if it has full access to the conditional probability distribution $p(r|x)$, by essentially inverting the encoding process i.e. by considering $p(r|x)$ as a function of the stimulus variable x. This constitutes the *likelihood function*. The *maximum likelihood decoder*, i.e. the decoder that selects the estimate $\hat{x}$ with highest likelihood, is a popular decoder under conditions of minimal assumptions.

More powerful, however, is a decoder that also takes into account that different values of x do not necessarily occur with same probability, but rather follow some ''prior'' probability distribution $p(x)$. If the decoder knows (or believes to know) this distribution, it could refine an estimate it otherwise would have performed based on the likelihood function alone, by computing the probability of a particular value of the stimulus variable x given the observed response r, written as the conditional distribution $p(x|r)$. Bayes' identity $p(x|r) = 1/p(r)p(r|x)p(x)$ tells us that this conditional probability distribution (called the ''posterior'') is exactly given by the normalized product between the likelihood function $p(r|x)$ and the prior probability $p(x)$. Decoding is completed by choosing an appropriate value $\hat{x}$. Again, popular choices are estimates $\hat{x}$ that have maximal probability (MAP, *maximum a posteriori*), or reflect the posterior mean. In general, the specific choice depends on how the overall estimation errors are weighted (loss function). We refer to this decoder as the *Bayesian observer model,* which is a simple description of a rational observer that correctly combines sensory and prior information when performing an estimate.

Note that the Bayesian observer model, although it is typically referred to as a ''decoding model'', contains full information about the encoding process via the likelihood function.

The ''invariance challenge''

Because the same stimulus content (e.g. object identity or motion direction) can exist in the world under conditions that produce very different retinal images, the visual system must find a means of associating retinal images

that contain the same content (despite their differences). In contrast to the ambiguity challenge, which has largely been studied by psychophysicists, the invariance challenge has largely been addressed by the physiology and computer vision communities. One likely reason for the separation is that the encoding/decoding framework presented in Figure 1, and implicitly assumed by most psychophysical models, has not proven useful for describing how the visual system deals with invariance. To understand why, it is important to recognize that any encoding description that simply maps a stimulus variable x directly onto the responses of an invariant neural population r does not address the invariance challenge. For example, a model that describes a direct mapping from different moving patterns parameterized by their motion direction to the directionally selective responses of neurons in MT need not describe *how* motion direction was extracted from the specific spatiotemporal patterns of light. Thus any model that addresses the invariance challenge must include a description of how an estimate of a stimulus variable $\hat{x}$ is extracted from a retinal image $I(x)$. More subtly, some have argued that attempting to describe the highly complex mapping from $I(x)$ to the stimulus estimate $\hat{x}$ with a two-stage encoding/decoding framework may be possible in theory but not feasible in practice; one example is the highly nonlinear transformation of a light-intensity based retinal image into a representation of object identity invariant to changes in an object's position, size, background and pose [24,25••]. As described below, models that address the invariance challenge are often motivated by the multi-stage structure of the visual system and thus extend the encoding/decoding framework to include a cascade of gradual, simpler operations.

Neurons implicated in invariant representations are much better described as ''tolerant'' than ''invariant'' in that these neurons do not tend to perfectly maintain their firing rate responses to different stimuli across conditions (e.g. changes in position), rather, they tend to maintain their relative preferences for particular stimulus variables across these changes. Under this definition, solutions to several invariance problems have been identified in the responses of visual neurons: V1 complex cells respond to an oriented bar in a manner largely independent of whether it is bright or dark [26]; a subset of MT neurons are tuned for motion direction in a manner largely independent of the particular moving pattern [27]; a subset of V2 and V4 neurons signal the relative depths of two surfaces at least somewhat independent of absolute depth [28,29]; a subset of MT neurons maintain their preferences for the orientation of a 3-dimensional surface ('tilt') in a manner largely independent of changes in absolute depth [30]; V4 neurons tend to maintain their selectivity for curvature over a range spatial positions [31]; and many IT cells maintain their selectivity for objects across changes in an object's position and size [32,33]. Recent theoretical work supports the idea that within some limitations, these tolerances, measured in individual neurons, can combine to form neural populations that robustly support the estimation of a particular stimulus variable in the face of other variation [34•].

How does the visual system produce neural responses that preserve their relative selectivity for stimulus variables tolerant to other variation? Hubel and Wiesel's initial descriptions of V1 complex cells proposed that a complex cell might extract orientation invariant to bar polarity by combining input from simple cell subunits with the same orientation preference but different bar polarity sensitivities ([26], formalized in the ''Energy Model'' [35]). While the anatomical validity of this hypothesis continues to be a topic of debate [36], this formulation continues to be a useful functional (i.e. mathematical) description of the response properties in these cells (e.g. [37]). Recently, similar ideas have been extended to describe invariant computation in visual areas beyond V1. All of these models include an initial stage of processing that converts the retinal image into the responses of a V1 population, followed by an invariant computation in a higher visual area. For example, one recent reported accounted for the preservation of V4 curvature selectivity over changes in position using a model in which curvature-tuned units were first produced by combining V1 complex cells with different spatial position and orientation preferences. This was followed by the combination of units with similar curvature tuning but different position preferences to confer units with position-tolerant, curvature-tuned responses [38]. Another recent study captured the invariance of relative-depth tuned V2 neurons across changes in absolute depth with a model that combined V1 units tuned for absolute depth in a manner similar to the ''Energy Model'' description of a V1 complex cell [28]. Yet another recent report described how MT neurons extract motion direction invariant to the particular pattern of the moving stimulus (i.e. gratings versus plaids) using a cascaded framework [39•,40]. In the first stage of this model, spatiotemporal motion patterns were converted into the responses of a V1 directionally selective population. The second stage of the model applied a simple linear (excitatory and inhibitory) weighting profile to the V1 input, configured to produce a matched direction preference invariant to the particular moving pattern.

These models are consistent with the notion that each visual area may implement a similar ''canonical'' computation albeit upon different inputs to produce an increasingly complex representation as signals propagate along the visual system [41,42,43••,44]. Notably, many successful models that describe the transformation of pixel intensity into an invariant representation of object identity (e.g. across position, size, background and pose) incorporate a relatively simple, canonical, feed-forward

framework (e.g. [42,45–48]). Intuitively, each stage of processing in these models implements a small amount of increased ''selectivity'' for conjunctions of simpler features (e.g. tuning for bars becomes tuning for contours) as well as a small amount of ''invariance'' for other variations (e.g. in the specific position of the contour)[43••]. The end-product of these successive stages is a population representation (e.g. in IT) in which simple neural machinery (such as a linear weighting function) can be applied to the population representation to extract object identity despite other stimulus variation [25••]. Recent experiments verify that such representations exist in IT [49].

Addressing both ambiguity and invariance

Thus far, we have treated ambiguity and invariance as if they were distinct challenges, i.e. as if at any one moment, the visual system has to deal with either one or the other. In reality, it seems more likely that the visual system faces both challenges simultaneously: while the invariance challenge is essentially synonymous with the task of recognition, the visual information required for recognition is often incomplete or missing. Thus it is often advantageous for the visual system to use contextual, prior information when solving a recognition task. In fact, behavioral evidence suggests that the visual system consistently relies on prior information to improve recognition performance even under conditions in which the visual information alone is sufficient for recognition. For example, several studies have demonstrated that humans recognize objects more quickly when presented in an expected as compared to an unexpected context (see [50] for a review). Similarly, the prior expectation for slow visual speeds biases the percept of moving objects toward slower speeds even under high contrast conditions [9].

Given that the visual system simultaneously deals with ambiguity and invariance under natural viewing conditions, a complete understanding of visual processing requires models that incorporate solutions for both computational challenges. The simplest instantiation of this modeling effort would be to address invariance and ambiguity in two sequential steps, i.e. a multi-stage model that describes the conversion of $I(x)$ into an invariant representation r of the stimulus variable x, followed by a Bayesian observer model to explain how the brain considers prior information when arriving at an estimate $\hat{x}$. In such model, contextual information could only be applied to the highest level of representation (e.g. to the identity of a person but not to the low-level statistics of the world).

As a more sophisticated approach, ''Hierarchical Bayesian models'' extend the Bayesian observer model to incorporate prior information at each level of a multi-stage processing hierarchy [51]. Invariant recognition could be achieved with feed-forward computation in such a model where priors reflect the overall statistical distributions of the stimulus features represented at each level. For example, prior information about the distributions of orientation and spatial frequency in natural scenes could be incorporated into the tuning characteristics and distributions of neurons in area V1. The ''Efficient coding'' hypothesis suggests that the tuning of sensory neurons should be distributed in a manner that efficiently represents image statistics [52]. However, how an efficient representation of image statistics relates to a resolution of ambiguity in the Bayesian framework remains little understood.

In addition to feed-forward computation, feedback operations in these hierarchical Bayesian models can provide a means to incorporate statistical dependencies between stimulus representations at different levels of abstraction. For example, the task of identifying a person from a retinal image can be mapped to a hierarchical Bayesian model where the lowest level represents simple image features like edges, that are then combined into increasingly complex subparts, up to the highest level that represents identity [53–55]. The feedback connections can be understood as a prediction of the higher level representation for what is expected at lower levels (e.g. ''If the person is Suzie, then the eyes are likely to be brown''). The lower level then uses this prediction as prior information and combines it with input from an earlier stage [56]. Furthermore, feedback connections provide a way to model how the visual system could propagate perceptual decisions down the hierarchy, explaining, for example, some of the reported perceptual biases in estimating motion direction contingent on an earlier decision about motion category [57•].

In sum, Hierarchical Bayesian models provide a potential framework for understanding how the visual system simultaneously deals with ambiguity and invariance. Some recent studies have begun to explore how humans can learn [58••] and perform inference [59,60] in hierarchical representations. How these models might be implemented in the neural architecture of the visual system is an issue that we are just beginning to address [61•,62].

Conclusions

We began this discussion by simplifying visual processing as a two-stage encoding/decoding process. We have demonstrated how two fundamental challenges of visual processing, ambiguity and invariance, can be formulated in this framework and we have described how multi-stage extensions of this model might incorporate both prior knowledge to resolve ambiguity as well as invariance to extract stimulus content. Notably, the ambiguity problem has primarily been emphasized by the theoretical and psychophysical communities whereas the invariance problem has primarily been emphasized by the communities

studying computer vision and physiology. Future progress in our understanding of visual processing will undoubtedly benefit from discussions and collaborative efforts between these subfields.

References and recommended reading

Papers of particular interest published within the period of review have been highlighted as:

- • of special interest
- •• of outstanding interest

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

2. Kersten D, Yuille A: **Bayesian models of object perception**. *Curr Opin Neurobiol* 2003, **13**:1-9.

3. Helmholtz H: *Treatise on Physiological Optics*. Bristol, UK: Thoemmes Press; 2000. Originally published in German in 1867 titled: 'Handbuch der Physiologischen Optik'..

4. • Torralba A: **How many pixels make an image?** *Vis Neurosci* 2009, **26(1)**:123-131.
This paper provides convincing demonstrations of how the human visual system uses contextual and prior information to recognize objects. Provided with images from real-world scenes of drastically reduced resolution (32 × 32 pixels), human subjects were nonetheless capable of identifying an average of five objects in these images, despite the fact that some objects were effectively only represented by a single pixel.

5. Knill DC, Richards W (Eds): *Perception as Bayesian Inference*. Cambridge University Press; 1996.

6. Weiss Y, Simoncelli E, Adelson E: **Motion illusions as optimal percept**. *Nature Neurosci* 2002, **5(6)**:598-604.

7. Dong DW, Atick JJ: **Statistics of natural time-varying images**. *Network: Comput Neural Syst* 1995, **6**:345-358.

8. Thompson P: **Perceived rate of movement depends on contrast**. *Vision Res* 1982, **22**:377-380.

9. Stocker AA, Simoncelli EP: **Noise characteristics and prior expectations in human visual speed perception**. *Nat Neurosci* 2006, **9(4)**:578-585.

10. Brainard DH, Longere P, Delahunt PB, Freeman WT, Kraft JM, Xaio B: **Bayesian model of human color constancy**. *J Vision* 2006, **6**:p1267ff.

11. Mamassian P, Landy M, Maloney LT: **Probabilistic models of the brain**. In *Bradford Book*. Edited by Rao RP, Olshausen BA, Lewicki MS. Cambridge, MA, USA: MIT Press; 2002:13-36.

12. Ernst MO, Banks MS: **Humans integrate visual and haptic information in a statistically optimal fashion**. *Nature* 2002, **415**:p429ff.

13. Knill D: **Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant**. *J Vision* 2007, **7(7)**:1-24.

14. Girshick AR, Banks MS: **Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts**. *J Vision* 2009, **9(9)**:8 1–20.

15. Vincent BT, Baddeley RJ, Troscianko T, Gilchrist ID: **Optimal feature integration in visual search**. *J Vision* 2009, **9(5)**:15 1–11.

16. Kording K, Wolpert D: **Bayesian integration in sensorimotor learning**. *Nature* 2004, **427(15)**:244-247.

17. Deneve, SaL P, Pouget A: **Efficient computation and cue integration with noisy population codes**. *Nature Neurosci* 2001, **4(8)**:p826ff.

18. Jazayeri M, Movshon JA: **Optimal representation of sensory information by neural populations**. *Nat Neurosci* 2006, **9(5)**:690-696.

19. Ma WJ, Beck JM, Latham PE, Pouget A: **Bayesian inference with probabilistic population codes**. *Nature Neurosci* 2006, **9**:p1432ff.

20. Zemel RS, Dayan P, Pouget A: **Probabilistic interpretation of population codes**. *Neural Comput* 1998, **10**:403-430.

21. Rao RPN: **Bayesian computation in recurrent neural circuits**. *Neural Comput* 2004, **16(1)**:1-38.

22. Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A: **Probabilistic population codes for Bayesian decision making**. *Neuron* 2008, **60(6)**:1142-1152.

23. Stocker AA, Majaj N, Tailby C, Movhshon JA, Simoncelli EP: **Decoding stimulus velocity from population responses in area MT of the macaque**. *Conference Abstract: Comput Systems Neurosci*. 2010. doi:10.3389/conf.fnins.2010.03.00298.

24. Marr D: *Vision*. W.H. Freeman; 1986.

25. •• DiCarlo JJ, Cox DD: **Untangling invariant object recognition**. *Trends Cogn Sci* 2007, **11(8)**:333-341.
This paper provides an insightful and highly intuitive graphical description about what makes invariant object recognition so challenging as well as how the visual system might solve it via a hierarchical cascade of simple operations.

26. Hubel DH, Wiesel AN: **Receptive fields, binocular interaction and functional architecture in the cats visual cortex**. *J Physiol* 1962, **160**:106-154.

27. Movshon JA, Adelson EH, Gizzi MS, Newsome WT: **The analysis of moving visual patterns**. *Exp Brain Res* **(Suppl. 11)**:1985: 117-151.

28. Thomas OM, Cumming BG, Parker AJ: **A specialization for relative disparity in V2**. *Nat Neurosci* 2002, **5(5)**:472-478.

29. Umeda K, Tanabe S, Fujita I: **Representation of stereoscopic depth based on relative disparity in macaque area V4**. *J Neurophysiol* 2007, **98(1)**:241-252.

30. Nguyenkim JD, DeAngelis GC: **Disparity-based coding of three-dimensional surface orientation by macaque middle temporal neurons**. *J Neurosci* 2003, **23(18)**:7117-7128.

31. Pasupathy A, Connor CE: **Shape representation in area V4: position-specific tuning for boundary conformation**. *J Neurophysiol* 2001, **86(5)**:2505-2519.

32. Ito M, Tamura H, Fujita I, Tanaka K: **Size and position invariance of neuronal responses in monkey inferotemporal cortex**. *J Neurophysiol* 1995, **73(1)**:218-226.

33. Tovee MJ, Rolls ET, Azzopardi P: **Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque**. *J Neurophysiol* 1994, **72(3)**:1049-1060.

34. • Li N, Cox DD, Zoccolan D, DiCarlo JJ: **What response properties do individual neurons need to underlie position and clutter "invariant" object recognition?** *J Neurophysiol* 2009, **102(1)**:360-376.
Through a combined experimental and modeling approach, this paper explores the single-neuron response properties that successfully and unsuccessfully combine, as a population, to support invariant recognition.

35. Adelson EH, Bergen JR: **Spatiotemporal energy models for the perception of motion**. *J Opt Soc Am* 1985, **2(2)**:284-299.

36. Martinez LM, Alonso JM: **Complex receptive fields in primary visual cortex**. *Neuroscientist* 2003, **9(5)**:317-331.

37. Rust NC, Schwartz O, Movshon JA, Simoncelli EP: **Spatiotemporal elements of macaque v1 receptive fields**. *Neuron* 2005, **46(6)**:945-956.

38. Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T: **A model of V4 shape selectivity and invariance**. *J Neurophysiol* 2007, **98(3)**:1733-1750.

39. • Rust NC, Mante V, Simoncelli EP, Movshon JA: **How MT cells analyze the motion of visual patterns**. *Nat Neurosci* 2006, **9(11)**:1421-1431.

This paper fits models to individual MT neurons that describe how these cells compute the direction of motion invariant to the specific moving pattern. Invariance is achieved in this model by a simple weighting profile applied to the input from V1.

40. Simoncelli E, Heeger D: **A model of neuronal responses in visual area MT**. *Vision Res* 1998, **38(5)**:743-761.

41. Douglas RJ, Martin KAC, Whitteridge D: **A canonical microcircuit for neocortex**. *Neural Comput* 1989, **1**:480-488.

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

43. •• Kouh M, Poggio T: **A canonical neural circuit for cortical nonlinear operations**. *Neural Comput* 2008, **20(6)**:1427-1451.
This paper provides a singular canonical form for cortical computation that can be adjusted to describe both selectivity and invariance computations at each stage of cortical processing.

44. Heeger DJ, Simoncelli EP, Movshon JA: **Computational models of cortical visual processing**. *Proc Natl Acad Sci USA* 1996, **93**:623-627.

45. Riesenhuber M, Poggio T: **Hierarchical models of object recognition in cortex**. *Nat Neurosci* 1999, **2(11)**: 1019-1025.

46. Serre TL, Wolf S, Bileschi S, Riesenhuber M, Poggio T: **Robust object recognition with cortex-like mechanisms**. *IEEE Trans Pattern Anal Mach Intell* 2007, **29**:411-426.

47. LeCun, Y, F-J H., and L Bottou. **Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting**. in *Proceedings of CVPR'04*. 2004: IEEE Press.

48. Wyss R, Konig P, Verschure PF: **A model of the ventral visual system based on temporal stability and local memory**. *PLoS Biol* 2006, **4(5)**:pe120.

49. Hung CP, Kreiman G, Poggio T, DiCarlo JJ: **Fast readout of object identity from macaque inferior temporal cortex**. *Science* 2005, **310(5749)**:863-866.

50. Oliva A, Torralba A: **The role of context in object recognition**. *Trends Cogn Sci* 2007, **11(12)**:520-527.

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

52. Barlow HB: **Possible principles underlying the transformation of sensory messages**. In *Sensory Communication.* Edited by Rosenblith WA. Cambridge, MA: MIT Press; 1961:217-234.

53. Jin Y, Geman S: **Context and hierarchy in a probabilistic image model**. In *Proceedings of CVPR'06*. IEEE Press; 2006.

54. Ommer B, Buhmann JM: **Learning the compositional nature of visual object categories for recognition**. *IEEE Trans Pattern Anal Mach Intell* 2010, 32(3): 501–516.

55. Kokkinos I, Yuille A: **HOP: hierarchical object parsing**. In *Proceedings of CVPR'04*. IEEE Press; 2009.

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

57. • Stocker AA, Simoncelli EP: **A Bayesian model of conditioned perception**. *NIPS Advances in Neural Information Processessing Systems 2010*. Cambridge: MIT Press; 2008, p.1409-1416.
This paper proposes an observer model in which top-down feedback reflects a prior conditioned on a previous categorical decision of the observer. The model proposes that a categorical decision can lead to substantial biases in subsequent estimation tasks. The authors argue that this sub-optimal behavior reflects the observer's attempt to ensure that the estimate is in agreement with that decision. Recent perceptual data from a variety of perceptual and cognitive tasks support the model.

58. •• Kemp C, Tenenbaum JB: **The discovery of structural form**. *Proc Natl Acad Sci USA* 2008, **105(31)**:10687-10692.
This paper presents a computational approach for learning the inherent structures of data sets. Its important contribution is the demonstration that the proposed approach is capable of reproducing known structures of perceptual spaces (among others) that accurately reflect perceptual distances. Although presented for general classes of data, the method might prove important for learning hierarchical Bayesian models of higher level vision.

59. Orban G, Fiser J, Aslin RN, Lengyel M: **Bayesian learning of visual chunks by human observers**. *Proc Natl Acad Sci USA* 2008, **105(7)**:2745-2750.

60. Feldman J: **Bayes and the simplicity principle in perception**. *Psychol Rev* 2009, **116(4)**:875-887.

61. • Shi L, Griffiths TL: **Neural implementation of hierarchical Bayesian inference by importance sampling**. *NIPS Advances in Neural Information Processessing Systems 2009*. Cambridge: MIT Press; 2009.
An interesting proposal of how a hierarchical Bayesian model can be implemented in neural architecture. The authors suggest that inference along the hierarchy can be approximated with importance sampling, and implemented with the neurally plausible mechanisms of local summation and normalization.

62. Friston K: **Hierarchical models in the brain**. *PLoS Comput Biol* 2008, **4(11)**:pe1000211.