In: The Cognitive Neurosciences V. M. Gazzaniga & R. Mangun, Eds. MIT Press. (2013)

Perception as an Inference Problem

Bruno A. Olshausen

Abstract. Although the idea of thinking of perception as in inference problem goes back to Helmholtz, it is only recently that we have seen the emergence of neural models of perception that embrace this idea. Here I describe why inferential computations are necessary for perception, and how they go beyond traditional computational approaches based on deductive processes such as feature detection and classification. Neural models of perceptual inference rely heavily upon recurrent computation in which information propagates both within and between levels of representation in a bi-directional manner. The inferential framework shifts us away from thinking of 'receptive fields' and 'tuning' of individual neurons, and instead toward how populations of neurons interact via horizontal and top-down feedback connections to perform collective computations.

Introduction

One of the vexing mysteries facing neuroscientists in the study of perception is the plethora of intermediate-level sensory areas that lie between low level and high level representations. Why in visual cortex do we have a V2, V3, and V4, each containing a complete map of visual space, in addition to V1? Why S2 in addition to S1 in somatosensory cortex? Why the multiple belt fields surrounding A1 in auditory cortex?

A common explanation for this organization is that multiple stages of processing are needed to build progressively more complex or abstract representations of sensory input, beginning with neurons signaling patterns of activation among sensory receptors in lower levels and culminating with representations of entire objects or properties of the environment in higher areas. For example, numerous models of visual cortex propose that invariant representations of objects are built up through a hierarchical, feedforward processing architecture (Fukushima 1980; Riesenhuber & Poggio 1999; Wallis & Rolls 1997). Each stage is composed of separate populations of neurons that perform feature extraction and spatial pooling, with information flowing from one stage to the next, as shown in Figure 1. The idea here is that each successive stage learns progressively more complex features of the input that are built upon the features extracted in the previous stage. By pooling over spatial position at each stage, one also obtains progressively more tolerance to variations in the positions of features, culminating in object-selective responses at the top level that are invariant to variations in pose of an object. Such networks now form the basis of 'Deep Learning' models in machine learning and have has achieved unprecedented success on both image and speech recognition benchmarks.

Might such hierarchical, feedforward processing models provide insight into what is going on in the multiple areas of sensory cortex? I shall argue here that despite the

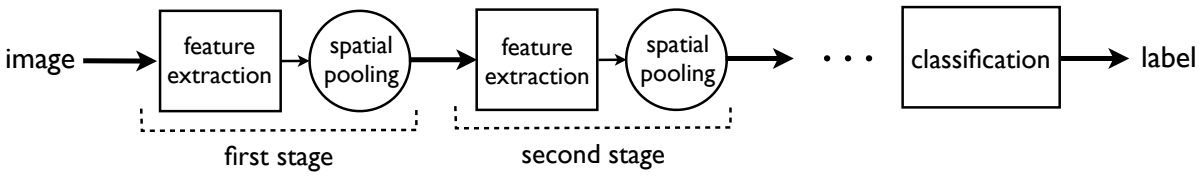

Figure 1: Hierarchical feedforward processing model of visual cortex. Each stage contains separate populations of neurons for feature extraction and pooling, resulting in object selective responses at the top stage that are invariant to variations in pose.

strong parallels between these models and cortical anatomy and physiology, these models are still missing something fundamental. The problem lies not just with their computational architecture, but also the class of problems they have been designed to solve. Namely, benchmark tasks such as image or speech recognition - while appearing to capture human perceptual capabilities - define the problem of perception too narrowly. Perception involves much more than a passive observer attaching labels to images or sounds. Arriving at the right computational framework for modeling perception requires that we consider the wider range of tasks that sensory systems evolved to solve.

So, what are these tasks? What do animals use their senses for? Answering these questions is a research problem in its own right. One thing we can say with certainty is that visual systems did not start out processing HD resolution images, and auditory systems did not start out with well-formed cochleas providing time-frequency analysis of sound. Rather, sensory systems began with crude, coarse-grained sensors attached to organisms moving about in the world. Visual systems for example began with simple light detectors situated in the epithelium. Remarkably, over a relatively short period of time (estimated to be 500,000 years) they evolved into the wide variety of sophisticated eye designs we see today (Nilsson & Pelger 1994). What was the fitness function driving this process? Presumably it was the ability to plan useful actions and predict their outcomes in complex, 3D environments. For this purpose, performance at tasks such as navigation or judging scene layout is crucially important. From an evolutionary perspective, the problem of 'recognition' - especially when distilled down to one of classification - may not be as fundamental it seems introspectively to us humans.

The greater problem faced by all animals is one of *scene analysis* (Lewicki, Olshausen, Surlykke & Moss, 2013). It is the problem of taking incoming sensory information and interpreting it in terms of what it conveys about the surrounding environment: terrain, obstacles, and navigable surfaces or routes, in addition to specific objects of interest and their pose and position within the scene. In contrast to a simple feedforward processing pipeline in service to the single goal of classification, scene analysis involves multiple types of representation with different functions (Figure 2). Most actions (the interesting ones we care about) are not simply reflexive behaviors in direct response to sensory messages. Rather, they depend on goals, behavioral state, and the past history of what has occurred (memory). In other words, meaningful behavior requires having a model of the world and one's place in it. The model needn't be particularly

detailed - indeed what aspects of the environment are modeled and to what degree of accuracy is an important empirical question - but we may reasonably assume that it must be assimilated from diverse types of sensory input into a common format that mediates planning and execution of behavior.

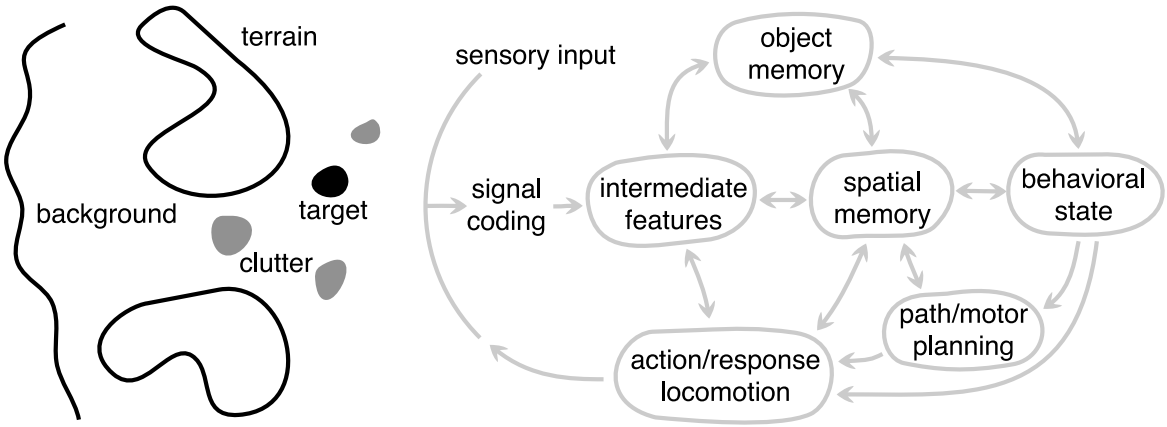

Figure 2: Components of scene analysis. The scene itself contains not just a single target object, but other objects, terrain, and background, all of which may be important for behavior. The neural structures enabling scene analysis contain multiple levels of representation and analysis. The level of "intermediate features" is where inferential processes come into play. (From Lewicki, Olshausen, Surlykke & Moss, 2013)

The goal of intermediate levels of representation then is to disentangle from the raw input stream aspects of the scene appropriate for driving behavior. In contrast to classification which collapses over variability to make a discrete categorical assignment, the goal here is to *describe* the variability - such as the slope of terrain, or the pose of an object - in an analog manner that captures the components of a scene and how one might act upon them. It is this intermediate level where inferential processes come into play.

In this chapter, I shall describe why inference provides a suitable computational framework for perception, the basic computations it entails, and specific models that have been proposed for how it is instantiated in neural systems. As we shall see, the inferential framework forces us to look at the neural mechanisms in a different way than the standard feedforward processing pipeline. An open problem though is to better understand the relation between perception and action, and how inferential computations fit into the larger framework of sensorimotor systems.

For other excellent reviews of 'perception as inference' from the perspective of psychophysics, see (Knill & Richards 1996; Kersten, Mamassian & Yuille 2004; Kersten & Yuille 2003; Yuille & Kersten 2006), in addition to the chapters by Kersten and Geisler in this volume.

Why Inference?

The problem of disentangling

The properties of the world that we care about - which drive behavior - are not directly provided by sensory input. There are no sensors that measure surface shape, motion of objects, material properties, or object identity. Rather, these properties are entangled among multiple sensor values and must be *disentangled* to be made explicit (see also DiCarlo & Cox 2007). In vision for example, the retinal image provides a set of measurements of how much light is impinging on the eye from each direction of space. The fact that we humans can look at 2D images and make sense of them unfortunately gives the misleading impression that an image tells you everything you need to know. But the image itself is simply a starting point and its 2D format is not well suited to drive behavior in a 3D world. Similarly, the array of hair cell responses does not provide an explicit representation of sound sources, nor does the array of mechanoreceptor activities on the fingertip provide a representation of object shape. These properties are entangled in spatio-temporal patterns of sensor activities.

Importantly, the nature of these disentangling problems is that they are often *ill-posed*, meaning that there is not enough information provided by the sensory data to uniquely recover the properties of interest. In other words, the various aspects of a scene that are needed to drive behavior can not simply be *deduced* from sensory measurements. Rather, they must be *inferred* by combining sensory data together with prior knowledge. Moreover, the disentangling often requires that different aspects of scene structure be estimated simultaneously, so that the inference of one variable affects the other. Thus, it would be impossible - or at least highly inefficient - to infer these things in a purely feedforward chain of processing.

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

In order to interpret this image, one must understand how illumination, 3D shape, and reflectance interact, and how an object combines with its background in projecting to a 2D image (i.e., occlusion). If an edge is ascribed to be due to a reflectance change, it can not also be due to a shape change (an edge could be due to both, but then the contribution of each of these causes would need to be reduced so that when combined they still match what is in the image). Thus, the computation of reflectance depends on the computation of shape, and vice-versa. And both of these require prior knowledge of

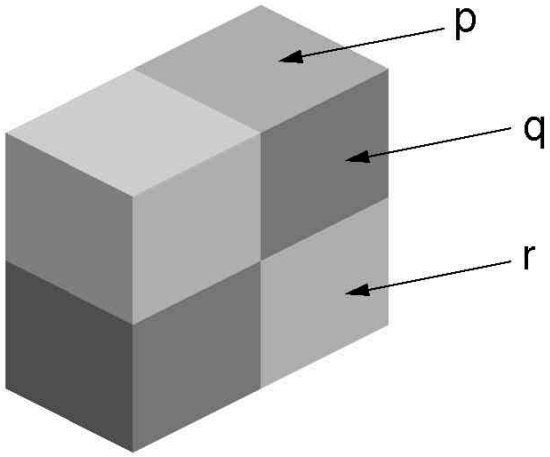

Figure 3: Image of a block painted in two shades of gray (from Adelson, 2000). The edges in this image are easy to extract, but understanding what they mean is far more difficult and can not be discerned through local image analysis.

what shapes and reflectance changes are likely in order to arrive at a plausible interpretation consistent with the data.

Early investigators such as Roberts and Waltz attempted to formally specify the logical operations needed to recover representations of 3D shape from such idealized "blocks world" scenes (Roberts 1965; Waltz 1975). However, their methods assumed perfect knowledge of edge segments in the image and the X and Y junctions formed at their intersections, and they utilized the constraints of geometry to then deduce 3D shape from the 2D image. Later, Marr (1982) proposed breaking this process into multiple stages: a primal sketch in which features and tokens are extracted from the image, a 2.5-D sketch that begins to make explicit aspects of depth and surface structure, and finally an object-centered, 3D model representation of objects. His model proposed a feedforward chain of processing in which features are extracted from the image and progressively built up into representations of objects through a logical series of operations in which information flows from one stage to the next. However, since these initial proposals, experience with real world images has shown us that such bottom-up, deductive processes rarely work in practice.

Consider for example the simple scene of a log against a background of rocks, as in Figure 4. It takes little conscious effort to comprehend what is going on in this scene - the boundary of the log appears obvious to most observers. But if we put ourselves in the position of a local population of neurons in V1 getting input from a local patch of this image, things are far less clear. The right panel of Figure 4 shows the response of an array of model V1, orientation-selective units analyzing a local patch of the image, with the boundary of the log superimposed as a faint gray line. As one can see, almost nowhere along this boundary are there neurons firing indicating the position and orientation of the boundary. Instead, one finds neurons firing at many different positions and orientations that signal structure in the background and foreground, but with little

relation to the boundary itself. Thus, simply measuring oriented contrast in an image does not give us a direct measure of the boundaries or intersection of objects, which can then be fed into a reasoning engine about 3D shape. The best that we will get from early levels of representation is a collection of ambiguous 2D shape cues which have aspects of illumination, shape, and reflectance intermingled. These must then must be aggregated and refined by higher level processes to disambiguate these cues and what aspects of scene structure they correspond to.

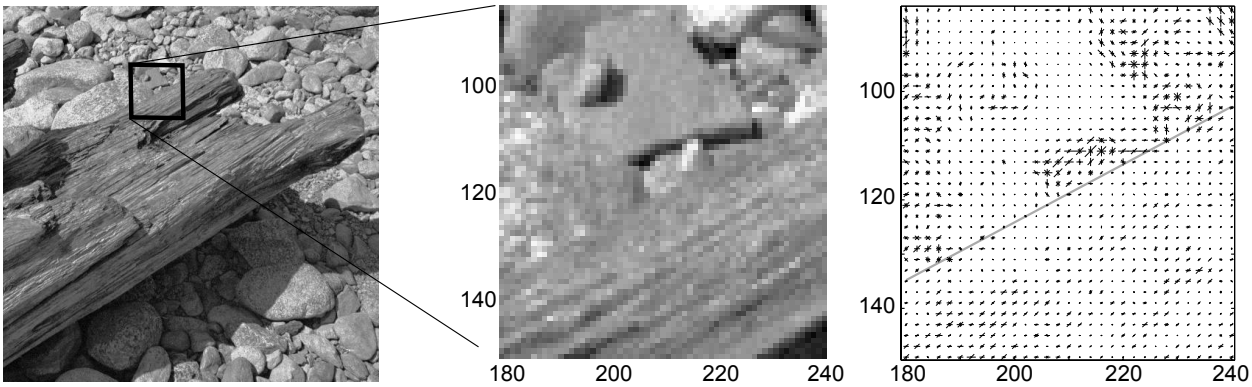

Figure 4. The outlined region around the boundary of the log (left panel) is shown expanded in the middle panel. The right panel shows how a hypothetical array of model V1 neurons (Gabor filters at four different orientations) would respond to the image subregion shown at left. The length of each line segment indicates the magnitude of response of a neuron whose receptive is situated at that position and orientation. An array of such neurons provides only weak or ambiguous cues about the presence of object boundaries in natural scenes.

‘Intrinsic images’

One of the first attempts to grapple with the computational aspects of the disentangling problem was Barrow and Tenenbaum’s work on ‘intrinsic images’ (Barrow & Tenenbaum 1978). They attempted to specify the rules by which scene components such as surface shape, reflectance, and illumination could be recovered from the raw intensity image. They argued that these attributes should be separated at an early level of representation, and that by doing so it greatly facilitates the process of segmentation and object recognition. Specifically, they proposed representing these attributes as a stack of two-dimensional maps, or ‘intrinsic images’, that are in register with the original intensity image, where each pixel location is labeled according to its shape, reflectance, or illumination properties. Importantly, the computation of each of these maps involves propagating information between maps to obey photometric constraints and within maps to obey continuity and occlusion constraints. That is, they can not be computed in independent streams in a purely feedforward fashion, but must cooperate to reach a solution.

Although Barrow and Tenenbaum appreciated the importance of disentangling and outlined some of the computational problems that need to be solved, they stopped short of proposing a specific algorithm and testing it on real images. Somewhat surprisingly,

the intervening years have seen only a handful of efforts to devoted to these problems (e.g., Jojic & Frey 2001; Wang & Adelson 1994) and as a result there has been little progress in developing practical solutions for dealing with real world images. Recently however Barron and Malik (2012) have made an important advance by using priors over shape, reflectance, and illumination to recover intrinsic images for these quantities from photographs of real objects. To date their method obtains the best performance on this challenging problem. And notably, it is based on inferential computation in which representations of shape, reflectance and illumination interact in order to settle to a solution.

The intrinsic image approach takes an important step in introducing the idea of a structured or layered representation that moves away from a flat, monolithic representation of image properties (such as an array of Gabor filters) and towards a representation of properties of the scene. But still, attributes of the scene are represented in 2D cartesian coordinates, in a retinotopic or camera-centric frame of reference, whereas animals must act in complex 3D environments. Ultimately then it makes sense for scene attributes to be represented in a format that is more amenable to planning actions in the world.

Surface representation

Nakayama and colleagues have argued based on psychophysical evidence that intermediate-level representations are organized around *surfaces* in the 3D environment, and that these representations serve as a basis for high-level processes such as visual search and attention (Nakayama, He & Shimojo 1995). This view stands in contrast to previous theories of perceptual grouping, search and attention based on 2D maps of image features such as local orientation and motion energy (Julesz 1981; Treisman & Gelade 1980). Nakayama's experiments suggest that representations of 3D surface structure are formed at an early stage, and that perceptual grouping, search and attention operate primarily on inferred surface representations rather than 2D maps of image features. For example, when colored items are arranged on surfaces in different depth planes, detection of an odd-colored target is facilitated when pre-cued to the depth plane containing the target; but if the items are arranged so as to appear attached to a common surface receding in depth, then pre-cueing to a specific depth has little effect. Thus, it would appear that attention spreads within inferred surfaces in 3D coordinates in the environment, not by 2D proximity in the image or within a common depth plane (disparity).

Nakayama's work also points to the importance of surface occlusion relationships in determining how features group within a scene. Under natural viewing conditions the 2D image arises from the projection of 3D surfaces in the environment. When these surfaces overlap in the projection, the one nearest the observer "over-writes" or occludes the other. Thus, a proper grouping of features would need to take this aspect of scene composition into account in determining what goes together with what, as shown in Figure 5. By manipulating disparity cues so as to reverse figure-ground relationships in a scene, they show that the visual system groups features in a way that

obeys the rules of 3D scene composition. Features are grouped within surfaces, even when parts of the surface are not visible, but not beyond the boundary of a surface. Thus, the neural machinery mediating this grouping would seem to require an explicit representation of border ownership, such as described by Zhou, Friedman and von der Heydt (2000), or some other variable that expresses the boundaries and ordinal relationship of surfaces.

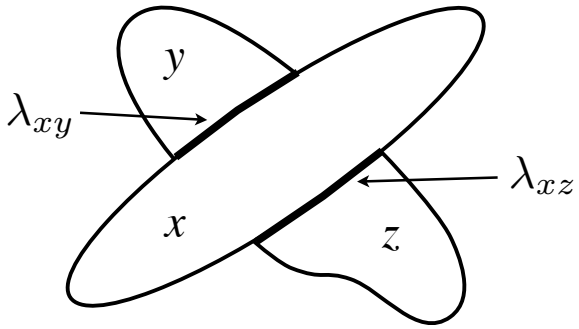

Figure 5: Occlusion and border ownership. When image regions corresponding to different surfaces meet in the projection of a scene, the region corresponding to the surface in front "owns" the border between them. A region that does not own a border is essentially unbounded and can group together with other unbounded regions. Here, surface x owns the borders λ_{xy} and λ_{xz}. Thus, regions y and z are unbounded at these borders and they are free to group with each other, but not with region x because it owns these borders and is therefore bounded by them. (Adapted from Nakayama et al. 1995)

Although the discussion above has focussed mainly on visual inferential processes, the same ideas generalize to other sensory modalities such as audition or touch. The central problem we face for all of these modalities is that the properties of the world that are needed to drive behavior are not given directly by the sensory receptors but instead must be inferred by combining sensory data together with prior knowledge. Now we turn to the question of how this is actually done by neurons.

How do neurons perform inferential computations?

Helmholtz astutely observed long ago that perception is a process of 'unconscious inferences.' Only recently though have investigators pursued this idea in a quantitative manner in order to characterize inferential computations carried out in the nervous system. Here I describe the mathematical framework for inference based on Bayes's rule, and neural models that have been proposed for doing perceptual inference.

Bayes's rule

The basic mathematical framework for inference begins with Bayes's rule, which uses the laws of conditional probability to calculate the probability of a hypothesis *H* given the data *D*:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

What this equation tells us is that if we have a model that specifies how probable the data would be under a certain hypothesis, i.e., the *likelihood* $P(D|H)$, in addition to the *prior* ('before data') probability of the hypothesis, $P(H)$, then we can calculate the *posterior* ('after data') probability of the hypothesis $P(H|D)$. (The term $P(D)$ often plays the role of a normalization constant and may be ignored if we are mainly interested in the relative probability of different hypotheses for the same data.) Simply speaking, Bayes's rule provides a calculus for reasoning in the face of uncertainty. It is a powerful conceptual and mathematical framework that tells us *quantitatively* how to make inferences in the face of noisy or incomplete data. Not surprisingly one now finds it applied to a wide variety of problems, from the control of guided missiles to spam filtering.

In perception, we are interested in estimating properties of the external environment from sensory data. For example, in vision we are given a set of photoreceptor activations or pixel intensities, I, and we wish to infer properties such as shape, s, and reflectance, r. Using Bayes's rule we can formulate this problem as follows:

$$P(s, r|I) \propto P(I|s, r)P(s)P(r)$$

Here the likelihood term $P(I|s, r)$ expresses the rendering model - i.e., how images are generated as a function of shape and reflectance. This is a well-posed computation that is routinely solved by computer graphics algorithms. However the problem of going the other direction - from the image to compute shape and reflectance - is highly ill-posed because there are multiple ways to set these parameters that would result in the same image (Adelson, 2000). This degeneracy is resolved by the priors over shape and reflectance, $P(s)$ and $P(r)$, which favor certain settings of s and r over others. In the work of Barron and Malik, these priors were obtained by measuring statistics of shape and reflectance on a large database of objects. The resulting posterior distribution over s and r, $P(s, r|I)$ rates the different shape and reflectance values of the image in terms of their probability of being the correct interpretation. It takes into account both how well image measurements are fit by these values and how consistent they are with prior knowledge. With strong enough priors, the posterior may be peaked around a single value of s and r, which would make these settings an obvious choice (i.e., the maximum-a-posteriori or MAP estimate).

A model that shows how the above computations may be implemented in a neurally plausible manner in the circuits of visual cortex has yet to be fully developed. In the

meantime though, we can get a feel for the nature of such a solution by looking at a simpler neural model of inference called *sparse coding*.

Sparse coding

The goal of sparse coding is to learn a set of basis patterns from the statistics of incoming sensory data and then infer a representation of the data in terms of these patterns. Although these patterns may not correspond to the actual properties of a scene, the model nevertheless illustrates the principles of inferential computation in a neural system and how it can provide new insight into the response properties of neurons.

In a visual sparse coding model, we start with the assumption that the spatial distribution of light intensities within a local region of the image $I(\vec{x})$ may be represented in terms of a superposition of some basis patterns $\phi(\vec{x})$:

$$I(\vec{x}) = \sum_i a_i \, \phi_i(\vec{x}) + n(\vec{x})$$

The image is then represented in terms of the coefficients a_i which tell us which basis patterns are contained in the image (Figure 6). The term $n(\vec{x})$ is a residual that is included to account for other structure such as noise that is not well described by the model. The basis patterns themselves are learned from the statistics of images so as to provide a sparse or compact description of the image - that is, we desire a *dictionary* of

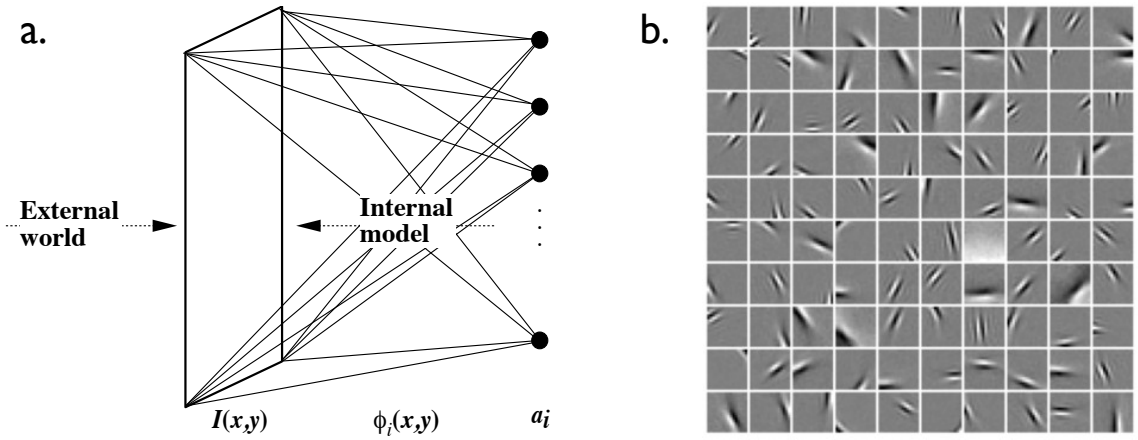

Figure 6: Sparse coding model of images. a. Images are represented in terms of a set of basis patterns $\{\phi_i(\vec{x})\}$ that are learned from the data. Each coefficient a_i expresses how much of each basis pattern is needed to describe the image. Sparsity is imposed through a prior that encourages coefficients to be zero. b. Basis patterns learned from image patches extracted from natural images. Each patch shows a different learned pattern $\phi_i(\vec{x})$. The learned patterns are oriented, localized, and bandpass (selective to structure at different spatial scales), similar to the measured receptive fields of V1 neurons.

basis patterns $\{\phi_i(\vec{x})\}$ that allows us to provide a good match to any given image using the fewest (sparse) number of non-zero coefficients a_i. Sparsity is enforced by imposing a prior over the coefficients $P(a)$ that encourages values to be zero. The coefficients themselves are then computed by maximizing the posterior distribution $P(a|I) \propto P(I|a)P(a)$.

The proposal here is that neurons in cortex (layer 4 of V1) are representing the coefficients a_i (Olshausen & Field 1997). But how should such a population of neurons compute their responses so as to maximize the posterior $P(a|I)$? Rozell, Johnson, Baraniuk and Olshausen (2008) have shown that a solution may be computed according to the following equations:

$$\begin{aligned} \tau \dot{u}_i + u_i &= \sum_{\vec{x}} \phi_i(x)\, I(\vec{x}) - \sum_j G_{ij}\, a_j \\ a_i &= g(u_i) \end{aligned}$$

These equations are amenable to direct implementation in a neural network (in fact they are the same equations as for a Hopfield network). Each neuron's membrane voltage u_i is driven by the combination of a feedforward (receptive field) term and a feedback (recurrent inhibition) term that depends on the overlap between basis patterns, $G_{ij} = \sum_{\vec{x}} \phi_i(\vec{x})\, \phi_j(\vec{x})$. The output a_i is then computed by simply thresholding the membrane voltage u_i (the function g passes values above a specified threshold and sets values below threshold to zero). Thus, in order to arrive at a representation of the image, the population of neurons must interact. Although the receptive field provides a driving input, the neuron's actual response is determined by the context of which other neurons around it are also responding. If the basis pattern of one unit is better matched to the image than another, it will attempt to cancel out or "explain away" the other unit's activity. Interestingly, Zhu and Rozell (2013) have shown that these explaining away interactions can account for a wide variety of non-classical receptive field effects such as end-stopping and contrast orientation tuning.

The sparse coding model illustrates how a variety of neural response properties found in V1 - localized, oriented, bandpass receptive fields and contextual modulation - may be accounted for in terms of a model that attempts to infer a representation of the incoming sensory data in terms of its underlying features. But the most we can hope for with this approach is a direct representation of the data per se (e.g., basis patterns of the image), whereas what we ultimately desire is a representation of the properties of a scene that these data tell us about. As the simple example of the painted block in figure 3 shows us, this can not be accomplished through local image analysis but rather involves aggregating information globally across the scene in order to infer properties such as shape and reflectance. Thus, we turn now to the question we addressed at the

outset: how can we build up more complex or abstract representations of sensory input through hierarchy of multiple stages of analysis?

Hierarchical representation

Lee and Mumford (2003) have proposed a framework for hierarchical Bayesian inference that illustrates how the above inferential computations could be extended to multiple stages of representation. The general idea is illustrated in Figure 7 which is adapted from their paper. At each stage, the variables being represented are influenced by both bottom-up and top-down inputs. At the first stage corresponding to V1, the variables *a* are inferred through a combination of the likelihood and prior as above, except now the prior over *a* is shaped by the variables *b* represented in the next higher level ('V2'). For example, if many weak signals among the *a* variables suggest the presence of a contour (as in the log and rocks image of figure 4), then the *b* variables in the next stage that explicitly represent the contour would become active, in turn encouraging the elements consistent with it to increase their activity by modulating the prior over *a* in this direction. The variables *b* in turn are subject to influences from yet higher levels, such as objects or fragments of surfaces represented by variables *c*. Thus, the full representation of the scene involves variables at all levels, *a*, *b*, and *c*, and computing these variables relies upon bi-directional information flow between levels.

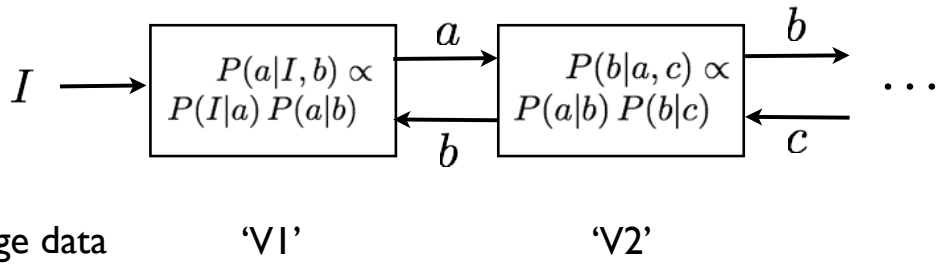

Figure 7. Hierarchical Bayesian inference. The variables represented at each level are inferred from a combination of bottom-up and top-down inputs. Bottom-up inputs enter into the likelihood, while top-down inputs enter into the prior. The two are combined to form the un-normalized posterior, which guides the inference of variables at each level. (Adapted from Lee and Mumford, 2003.)

This is admittedly just the sketch of a theory. There are many details to be filled in here, and some efforts have already been made along these lines (Cadieu & Olshausen 2012; Garrigues & Olshausen 2010; Karklin and Lewicki 2005, 2009). What is most needed now is to incorporate layered representations, such as those proposed by Barrow and Tenenbaum, and Barron and Malik, that separate aspects of scene structure due to surface shape, reflectance, or other scene variables. It will also be necessary to include in the generative model the ability to account for occlusion or figure-ground relationships (Le Roux, Heess, Shotton & Winn 2011; Lücke, Turner, Sahani & Henniges 2009) and geometric transformations due to variations in pose (Arathorn 2002, 2005; Olshausen, Anderson & Van Essen 1993).

From a neurobiological perspective, the hierarchical inference model provides a clear role for feedback connections in the cortex and it suggests how to design experiments to reveal what they are doing. Although there have already been numerous experimental attempts to uncover what feedback is doing - for example by cooling or disabling neurons in a higher area and characterizing how responses in lower areas change (Andolina, Jones, Wang & Sillito 2007; Angelucci & Bullier 2003; Hupé, James, Girard, Lomber, Payne & Buillier 2001) - the effects to date appear rather subtle. Indeed there is considerable doubt among neuroscientists as to whether feedback plays any role in dynamically shaping information processing (Lennie 1998). If the hierarchical Bayesian inference model is correct, it suggests we would most see the greatest effects of feedback when the system is presented with scenes containing locally ambiguous cues which can only be properly interpreted at higher levels of analysis. Indeed, fMRI experiments along these lines have revealed evidence for strong top-down effects: When subjects perceive a collection of features as an entire 3D object as opposed to its individual parts, activity in higher levels increases while activity in lower levels decreases, consistent with disambiguation (Murray, Kersten, Olshausen, Schrater & Woods 2002).

Another important neurobiological consideration is the speed with which such a hierarchical inference system can settle on a solution. It has been argued based on the speed of object-selective neural responses (Thorpe, Fize & Marlot 1996; Hung, Kreiman, Poggio & DiCarlo 2005; Oram & Pettet 1992) that there is little time for the iterative type of processing that feedback loops would entail (Thorpe & Imbert 1989). But conduction velocities of feedforward and feedback axons between V1 and V2 are on the order of 2-4 ms (Angelucci & Bullier 2003). Even between thalamus and V1 the round trip travel time can be as short as 9 ms (Briggs & Usrey 2007). It also not clear whether “iterative processing” is an apt analogy to describe signal flow in the cortex, since there is no clock latching each cycle of feedback. It could well be then that such a system can quickly settle to a solution within physiological time constraints. More importantly, sensory processing does not work in terms of static snapshots of input that get churned away in the system one at a time. Rather it operates as a dynamical system operating on a continuous, time-varying input stream. Thus, the consequence of feedback arriving through axonal and synaptic delays is simply that sensory information arriving at the present moment is processed in the context of past information that has gone through a higher level of processing, which undoubtedly could be quite advantageous.

Finally, it should be noted that the hierarchical Bayesian inference framework is distinct from the ‘predictive coding’ model of Rao and Ballard (1999). Though both models advocate an important role for top-down feedback signals, the proposed effect of these signals is very different. Predictive coding proposes that feedback signals are largely inhibitory, as they carry the predictions of higher levels which attempt to cancel out signals coming from lower levels. By contrast, Hierarchical Bayesian inference proposes that feedback serves to disambiguate representations in lower levels, meaning that it would facilitate, rather than cancel out, the activity of neurons at lower

levels consistent with representations from higher levels, and it would suppress the activity of neurons that are inconsistent.

Conclusions

Thinking of perception as an inference problem, as opposed to a deductive computational pipeline, leads us to ask a different set of questions about what neurons are doing. Instead of asking about feature extraction, receptive fields, and tuning of individual neurons, we are led to ask how populations of neurons cooperate and interact to infer representations of scene properties. Instead of looking for organized "maps" of sensory features varying along one or a small number of feature dimensions across the cortical surface, we are led to look for different layers of representation which disentangle different scene properties and which are likely to be intermingled on a much finer scale. Instead of viewing the cortical hierarchy as a feedforward pipeline with classification as an end goal, we are led to ask how ill-posed problems are being solved at each level, and how feedback from higher levels disambiguates representations at lower levels.

While Bayes's rule provides a computational framework for perceptual inference, it leaves many questions unanswered regarding its implementation. If the brain is doing Bayesian inference, should we expect to find neurons representing probabilities and calculating Bayes's rule when we look inside? As we have seen in the case of the sparse coding model, not necessarily. Here neurons represent hypotheses about what is contained in the scene, and Bayes's rule acts as the 'invisible hand' that governs the dynamics of the circuit and leads the population to a mode of the posterior distribution. Other models have proposed ways that neurons might represent probabilities implicitly through a population code (Eliasmith & Anderson 2004; Ma, Beck, Latham & Pouget 2006) or through stochastic sampling via spontaneous activity (Berkes, Orban, Lengyel & Fiser 2011). The advantage of these latter approaches is that the entire distribution over a set of hypotheses may be represented which allows for uncertainty or multiple probable hypotheses to be taken into account during inference.

Finally, it will be necessary to think more seriously about the link between the perception and action in order to give these ideas a more solid footing. Much has been said here about representing scene properties, but which properties need to be represented and how well depends on the manner in which they are used to guide actions. As Guillery and Sherman (2011) point out, layer 5 neurons in all levels of visual cortex (or other sensory cortices) project to motor nuclei. Thus it is not just the top box of the cortical hierarchy but also representations in V1, V2, V4, etc. that are used to guide actions. Figuring out how to incorporate these aspects of cortical architecture into the hierarchical inference framework - that is, understanding how inference feeds into action - will be an important goal for future work.

Acknowledgments

These ideas evolved in large part through discussions with Mike Lewicki, Cindy Moss, and Anne-Marie Surlykke while on sabbatical at the Wissenschaftskolleg zu Berlin in 2009. I also thank Jim DiCarlo for encouraging me to write these things down and better explain the inferential framework of perception. Supported by NSF (IIS-1111654), NIH (EY019965), NGA (HM1582-08-1-0007), SRC STARnet (SONIC), and the Canadian Institute for Advanced Research.

References

Adelson, E. H. (2000). Lightness Perception and Lightness Illusions. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences, 2nd Ed.* (pp. 339-351): MIT Press.

Andolina, I. M., Jones, H. E., Wang, W., & Sillito, A. M. (2007). Corticothalamic feedback enhances stimulus response precision in the visual system. *Proceedings of the National Academy of Sciences of the United States of America, 104*(5), 1685-1690. doi: 10.1073/pnas.0609318104

Angelucci, A., & Bullier, J. (2003). Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? *Journal of Physiology - Paris, 97*(2-3), 141-154. doi: 10.1016/j.jphysparis.2003.09.001

Arathorn, D. W. (2002). *Map-seeking circuits in visual cognition: A computational mechanism for biological and machine vision*: Stanford University Press.

Arathorn, D. W. (2005). Computation in the Higher Visual Cortices: Map-Seeking Circuit Theory and Application to Machine Vision. *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop (AIPR'04)*, 1-6.

Barron, J. T., & Malik, J. (2012). Shape, Albedo, and Illumination from a Single Image of an Unknown Object. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barrow, H. G., & Tenenbaum, J. M. (1978). Recovering Intrinsic Scene Characteristics from Images. In A. Hanson & E. Riseman (Eds.), *Computer Vision Systems* (pp. 3-26): Academic Pr.

Berkes, P., Orban, G., Lengyel, M., & Fiser, J. (2011). Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science, 331*(6013), 83-87. doi: 10.1126/science.1195870

Briggs, F., & Usrey, W. M. (2007). A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey. *Journal of Neuroscience, 27*(20), 5431-5436. doi: 10.1523/JNEUROSCI.1035-07.2007

Cadieu, C. F., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation, 24*(4), 827-866. doi: 10.1162/NECO_a_00247

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333-341. doi: 10.1016/j.tics.2007.06.010

Eliasmith, C., & Anderson, C. (2004). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*: MIT Press.

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

Garrigues, P. J., & Olshausen, B. A. (2010). Group sparse coding with a laplacian scale mixture prior. *Advances in Neural Information Processing Systems, 24*.

Guillery, R. W., & Sherman, S. M. (2011). Branched thalamic afferents: What are the messages that they relay to the cortex? *Brain Research Reviews, 66*(1-2), 205-219. doi: 10.1016/j.brainresrev.2010.08.001

Hung, C. P., Kreiman, G., Poggio, T., & Dicarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science, 310*(5749), 863-866. doi: 10.1126/science.1117593

Hupe, J.-M., James, A. C., Girard, P., Lomber, S. G., Payne, B. R., & Bullier, J. (2001). Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology, 85*(1), 134-145.

Jojic, N., & Frey, B. J. (2001). *Learning flexible sprites in video layers*. Paper presented at the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=990476

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290*(5802), 91-97. doi: 10.1038/290091a0

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation, 17*(2), 397-423.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*(7225), 83-U85. doi: 10.1038/nature07481

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. doi: 10.1146/annurev.psych.55.090902.142005

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology, 13*(2), 150-158.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*: Cambridge University Press.

Le Roux, N., Heess, N., Shotton, J., & Winn, J. (2011). Learning a generative model of images by factoring appearance and shape. *Neural Computation, 23*(3), 593-650. doi: 10.1162/NECO_a_00086

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, A, 20*(7), 1434-1448.

Lennie, P. (1998). Single units and visual cortical organization. [Review]. *Perception, 27*(8), 889-935.

Lewicki, M. S., Olshausen, B. A., Surlykke, A., & Moss, C. F. (2013) Scene analysis in the natural environment. (submitted)

Lücke, J., Turner, R. E., Sahani, M., & Henniges, M. (2009). *Occlusive components analysis*. Paper presented at the Advances in Neural Information Processing Systems. http://eprints.pascal-network.org/archive/00007969/

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432-1438. doi: 10.1038/nn1790

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*: WH Freeman.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America, 99*(23), 15164-15169. doi: 10.1073/pnas.192579399

Nakayama, K., He, Z., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision *An invitation to cognitive science: Visual cognition* (Vol. 2, pp. 1-70): MIT Press.

Nilsson, D. E., & Pelger, S. (1994). A pessimistic estimate of the time required for an eye to evolve. *Proceedings Biological sciences / The Royal Society, 256*(1345), 53-58. doi: 10.1098/rspb.1994.0048

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience, 13*(11), 4700-4719.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research, 37*(23), 3311-3325.

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68*(1), 70-84.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87. doi: 10.1038/4580

Rao, R. P. N., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain: Perception and neural function*: MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025. doi: 10.1038/14819

Roberts, L. G. (1965). Machine Perception of Three-Dimensional Solids. In J. T. Tippett (Ed.), *Optical and Electro-optical Information Processing* (pp. 159-198): MIT Press.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. [Letter]. *Neural Computation, 20*(10), 2526-2563. doi: 10.1162/neco.2008.03-07-486

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522. doi: 10.1038/381520a0

Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist models. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulie & l. Steels (Eds.), *Connectionism in Perspective* (pp. 63-92): Elsevier Inc.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136. doi: 10.1016/0010-0285(80)90005-5

Waltz, D. (1975). Understanding Line Drawings of Scenes with Shadows. In P. H. Winston (Ed.), *The Psychology of Computer Vision*.

Wallis, G., & Rolls, E. T. (1997) Invariant face and object recognition in the visual system. Progress in Neurobiology, 51, 2, 167-194.

Wang, J. Y. A., & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions On Image Processing, 3*(5), 625-638. doi: 10.1109/83.334981

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308. doi: 10.1016/j.tics.2006.05.002

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience, 20*(17), 6594-6611.

Zhu, M., & Rozell, C. J. (2013). Visual Nonclassical Receptive Field Effects Emerge from Sparse Coding in a Dynamical System. *PLoS Comput Biol, 9*(8), (in press). doi: 10.1371/journal.pcbi.1003191