2 Vision, Psychophysics and Bayes

Paul Schrater and Daniel Kersten

Introduction

Neural information processing research has progressed in two often divergent directions. On the one hand, computational neuroscience has advanced our understanding of the detailed computations of synapses, single neurons, and small networks (cf. [27]). Theories of this sort have benefited from the scientific interplay between testable predictions and experimental results from real neural systems. As such, this direction is aimed most directly at the mechanisms implementing visual function.

On the other hand, theoretical neural networks have been subsumed as special cases of statistical inference (cf. [34, 3, 26]). This direction is well-suited to large scale systems modeling appropriate to the behavioral functions of perceptual and cognitive systems. We've seen considerable progress in the solution of complex problems of perception and cognition–solutions obtained without specific reference to neural or biological mechanisms. However, in contrast to small-scale neuron-directed modeling, behavioral theories face a different experimental challenge–namely, how can quantitative models be refined by experiment? The neural implementation of a model of statistical pattern recognition typically has too many independent variables to test neurophysiologically, and behavioral tests are unsatisfying because most of the parameters are unmeasureable.

The purpose of this chapter is to suggest that the development of neural information processing theories based on statistical inference is actually good thing for psychology, and in particular for visual psychophysics. The proper level of abstraction avoids the premature introduction of unmeasureable neural parameters that are too numerous or difficult to test behaviorally. Yet, as is the case for thermodynamics and models of molecular motion, the bridge between statistical pattern theories and neural networks can be made when required.

The principle concern of psychologists is behavior. In vision, we use the term

behavior broadly to include perceptual psychophysics, experimental studies of visual cognition, and visual motor control . One can also distinguish two often divergent directions in the study of visual behavior in questions that address: 1) neural mechanism and 2) functional tasks. These questions can often be rephrased in the form of questions that address what the visual system can do when pushed to extremes, in contrast to what it does do in natural circumstances.

What people can do. There is a long tradition of relating the phenomena of visual perception and psychophysical performance to underlying neural mechanisms. Probably the most famous and successful example is color trichromacy which was deduced from psychophysical experiments of Helmholtz, Maxwell and others in the 19th century. The neurophysiological basis in terms of discrete retinal receptor types wasn't firmly established until the middle of the 20th century. Another example is the intriguing correspondence between the wavelet-like patterns that humans detect best and the receptive field profiles of simple cells in V1 [7, 43, 20]. For tests of mechanism, the behavioral tasks can be rather unnatural and the stimuli often bear little resemblance to the kinds of images we typically see. Signal detection theory has played an important role in relating behavior to mechanism [15]. It provides a bridge between theory and experiment as applied to neural mechanisms. In this first case, statistical or signal detection theories provide the means to account for the information available in the task itself–a step often neglected in drawing conclusions about mechanism from psychophysical measurements.

What people do do. There is also an ecological tradition in which we seek to understand how visual function is adapted to the world in which we live. Answering the question "How well do we see the shapes, colors, locations, and identities of objects?" is an important component in providing answers to the more general question "How does vision work?". The challenge we address below is arriving at a quantitative model to answer the "How well do we see ..." question for natural visual tasks. Statistical inference theory, and in particular the specific form we describe below as pattern inference theory, plays a role in the analysis of both kinds of perceptual behavior–but is particularly relevant for developing quantitative predictive models of visual function.

In this chapter, we describe a framework within which to develop and test predictive quantitative theories of human visual behavior as pertains to both mechanism and function. What are we asking for in a quantitative theory? A complete quantitative account of a human visual task should include models both of function, and of the mechanisms to realize those functions.

Problems of vision: ambiguity and complexity

At the functional level, one would like a model whose input is a natural image or sequence and whose output, shows the same performance as the human along

some dimension. Further, the model should also predict the pattern of errors with respect to ground truth. Below we will describe a recent study which compared the perceived surface colors under quasi-natural lighting conditions with those of a Bayesian estimator.

Any such modeling effort immediately runs into two well-known *theoretical problems* of computer vision: ambiguity and complexity. How can a vision system draw reliable conclusions about the world from numerous locally ambiguous image measurements? The ambiguity arises because any image patch can be produced from many different combinations of scene variables (e.g. object and illumination properties). In addition, *all* of the scene variables, both relevant and irrelevant for the observer's task, contribute to natural image measurements. In general, image measurement space is high-dimensional, and building models that can deal with these problems leads to complex models that are difficult to analyze and for which it is difficult to determine what are the essential features. One possible solution is the historical one–take an empirical approach to human vision, and develop models to summarize the data from the ground up.

However, the complex high-dimensional nature of the relevant scene parameters leads to an *empirical problem* as well: psychophysical testing of models of human perception under natural conditions could rapidly become hopelessly complex. For example, with over a dozen cues to depth, a purely empirical approach to the problem of cue integration is combinatorially prohibitive. Thus, we are faced with two formidable problems in vision research. First, how do we develop relevant and analyzable models for vision, and second, how do we test these models experimentally?

Developing Theories

For the modeling problem, we will argue that it is crucial to specify the correct level of analysis, and we distinguish modeling at the functional level from modeling at the level of mechanisms. The functional level constitutes a description of the *information* required to perform for particular task without worrying about the specific details of the computations. It addresses issues of the observer's prior knowledge and assumptions about scene structure, image formation, and the costs associated with normal task demands. On the other hand, the mechanistic level is concerned with the specific details of neural computations.

Functional Level Modeling

At the functional level, we wish to model a natural visual task, like apple counting. Given the natural task, the solution we propose is to let the ideal (Bayesian) observer for the task serve as a default model. The modeling strategy is to hypothesize that human vision uses all the information optimally. Of course, human vision

is not optimal, but starting from the optimal observer yields a coherent research strategy in which models can be modified to discard the same kinds and amount of information as the human visual system. This approach departs from the historical bottom-up approach to modeling, but we suggest that it may be ultimately simpler than trying to determine how to "smarten up" a suboptimal model based on experiment. Here the main focus is how often the human is ideal or "ideal-like". By ideal-like we mean that peformance parallels ideal in all but a small number of dimensions (e.g. additive internal noise). Below we discuss an approach to information modeling we call "pattern inference theory", that is an elaboration of signal detection theory that provides the generative and decision theoretic tools to model the informational limits to natural tasks [22, 21, 25].

Two main strategies at the functional level can be distinguished: a) model the complexity of the visual inference, and compare the model with human performance, but don't try to directly model the information that is present in natural scenes [44]; b) model the physical information, and measure how well this model accounts for human performance. It is this second strategy we illustrate below with an example from color constancy [4]. Having a general purpose model for counting apples in trees may be in the distant future, but part of that model will involve understanding how surface colors for apples can be distinguished from those of leaves. Because of interreflections, any given point of the retinal image will often have wavelength contributions from direct and indirect light sources (i.e. from apples and leaves), that need to be discounted to infer the identity and properties of objects.

Mechanism Level Modeling

At the mechanism level we are interested in how visual stimuli are actually processed, what features in the image are measured, what kinds of representations does the visual system use. One would like an account of the neural systems, the spatial-temporal filtering, neural transformations and decision processes leading to the output. Here the modeling methods are more diffuse, encompassing optimal image encoding ideas [32, 47, 40] and more traditional analyses of the utility of image measurements for some task (e.g. the optic flow field for inferring egomotion).

In these kinds of study the model is typically not generated through an ideal observer analysis of a natural task. Nevertheless, we can profitably use the ideal observer approach to test mechanistic models. Here, deviations from ideal provide insight into biological limitations. The immediate focus of interest is how much human performance differs from ideal, because it is the differences which are diagnostic of the kinds of mechanisms used by the visual system. For example, we know that light discrimination departs from the ideal at high light levels, long durations, and large regions [2]. We also know that human pattern detection competes well with ideal observers when the patterns match the receptive field profiles

of simple cells [7, 43, 20]. The way in which spatial resolution departs from ideal can be explained in large part by optical, sampling, and neural inefficiencies [14].

In addition, having an understanding of the ideal observer for the task used in the laboratory to test the mechanistic hypotheses is crucial. Psychophysical testing hypotheses of mechanism leads to a problem in addition to that of complexity: inferring mechanisms from psychophysics must take into account how performance depends on the information for the task. In particular, *opposite conclusions can be drawn from psychophysical results depending on how the information is modeled*. (cf. [10, 23, 24, 9]) As we illustrate below, ideal observer analysis, a part of a signal detection theory, provides the solution to this problem of information normalization.

As an example of the difference between functional and mechanistic levels, consider the following question. What are the crucial features of leaf texture as distinct from apples that would lead to reliable segmentation? This kind of question can be addressed at the functional level by specifying generative texture models [47] for the task and testing whether the most statistically reliable features for discrimination are used by the human observer. However, one could imagine a finer grain analysis to test whether the particular features hypothesized by the information model are indeed processed in terms of particular spatial filters. At this level, laboratory manipulations can be geared towards analysing mechanisms of visual processing. In the context of our apple counting problem, part of a complete explanation will be to understand how motion parallax can help to segment apple from leaf surfaces–but underneath this level, is the question of the kind of motion mechanism human vision might use to support such functional inferences.

Testing models using ideal observers

One of the major points of this chapter is that the modeling problem naturally breaks into two: determining how a useful signal (e.g. objects, distances, shapes, etc) get *encoded* into intensity changes in the image, and second, determining the limits to *decoding* the image to infer the signals. The encoding problem, which involves modeling both the regularities and structure in the signal domain as well as the image formation process, has been frequently neglected in studies of human vision (but see [33]). We discuss this problem more completely in the section on Pattern Inference Theory below. On the other hand, the decoding problem involves finding decoding strategies that clearly depend on how the signals were encoded. In the study of decoding, the fundamental object is the optimal decoder, comparedto which all other decoders can be described in terms of the information about the signals they discard. Thus studying the decoding problem relies on theories of ideal observers, or more generally of optimal inference. If we describe human perception as a process of decoding images, then the ideal observer can be used to describe human performance in terms of deviations from optimality.

Returning to the question of how to test our models experimentally, the preceding suggests the strategy of comparing human to ideal performance. We will discuss how this comparison can be used to test theories and give two examples of such tests below.

In the next section, we outline the elements of a Bayesian approach to vision applicable to both the analysis of mechanism and function. In the following section, we apply these elements to questions of mechanism and in particular we illustrate information normalization in the analysis of motion measurements. The last section shows how information modeling has been used to provide a quantitative account of an aspect of functional vision, color matching results for simple natural images.

Bayesian perception and Pattern Inference Theory

Because observers use their visual systems to do things, theories of visual perception cannot be built in isolation from functional visual tasks. We see the consequences of this statement as including the following principles: 1) vision is inference; 2) vision has relevant knowledge of scene structure prior to performing a task; 3) the affordances (value of the outcomes) of a completed task determine the costs and benefits of acquiring visual knowledge. A fundamental approach that quantifies these principles is a theoretical apparatus we call *Pattern Inference theory*, which is a conjunction of Bayesian Decision Theory with *Pattern Theory*. As an elaboration of signal detection theory, we choose the words *pattern* and *inference* to stress the importance of modeling complex natural signals, and of considering tasks in addition to detection, respectively.

We have elsewhere [22, 46] argued that Pattern Inference Theory provides the best language for a quantitative theory of visual perception at the level of the naturally behaving (human) visual system. The term, "Pattern Theory" was coined by Ulf Grenander to describe the mathematical study of complex natural patterns [16, 17, 30, 46]. In our usage, Pattern Inference Theory is a probabilistic model of the observer's world and sensory input, which has two components: the objects of the theory, and the operations of the theory. The objects of the theory are the set of possible image measurements I, the set of possible scene descriptions S, and the joint probability distribution of S and I: $p(S, I)$. The operations are given by the probability calculus, with decisions modeled as cost functionals on probabilities. The richness of the theory lies in exploiting the structure induced in $p(S, I)$ by the regularities of the world (laws of physics) and by the habits of observers. This emphasizes a central role for the modeling of representations and their transformations. Pattern Inference theory also assumes a central role, in perception, of generative (or synthetic) models of image patterns, as well as prior probability

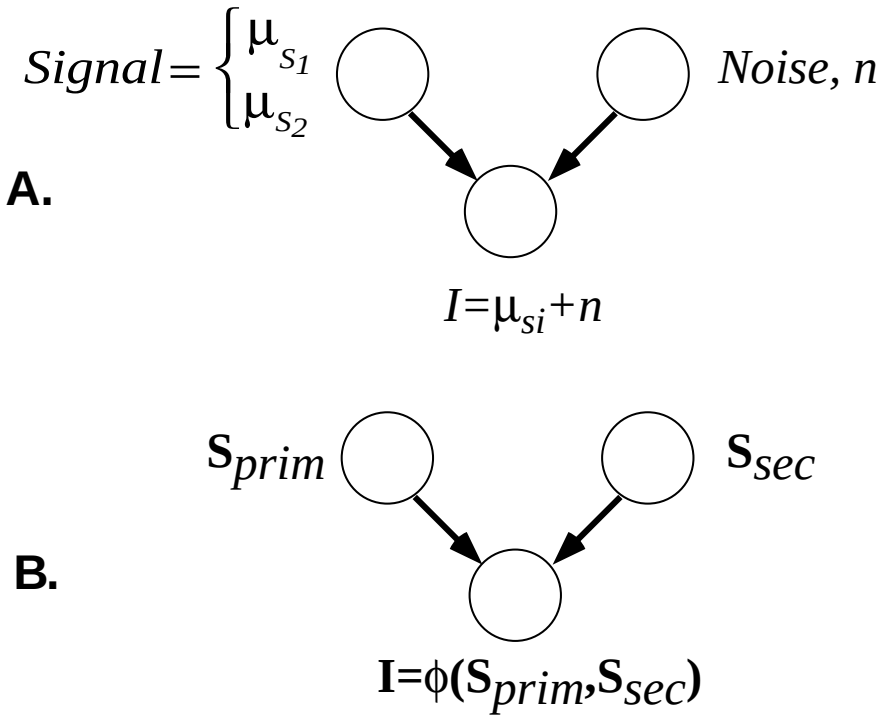

Figure 2.1: Panel A illustrates the generative model for the signal detection task. Signal detection theory provides tools particularly appropriate for the behavioral analysis of visual mechanisms. Panel B illustrates the general form of the generative model for the pattern inference theory task. Pattern inference theory is an elaboration of signal detection theory which seeks to take into account the specific generative aspects, $\phi()$ of natural image formation, and the full range of natural tasks. Pattern Inference Theory is of particular relevance for modeling natural visual function.

models of scene information. Our example below compares two generative models of color, one based on direct, and a second on direct plus indirect lighting. An emphasis on generative models, we believe, is essential because of the inherent complexity of the causal structure of high-dimensional image patterns. One must model how the multitude of variables, both needed and unneeded, interact to produce image data in order to understand how to decode those patterns.

How can we describe the processes of visual inference as image decoding by means of probability computations (i.e. from the point of view of pattern inference theory)? To do so requires a probabilistic model of tasks. We consider a task as specifying four ingredients: 1) the relevant or primary set of scene variables S_{prim}, 2) the irrelevant or secondary scene variables S_{sec}, 3) the scene variables which are presumed known S_f, and 4) the type of decision to be made. Each of the

four components of a task plays a role in determining the structure of the optimal inference computation[1].

Bayesian decision theory provides a precise language to model the costs of errors determined by the choice of visual task [45, 6]. The ideal observer that minimizes average error finds the $S_{prim}*$ which minimizes the following risk:

$$R(S_{prim}; I, S_f) = -\int_{S_{sec}*} P(S_{sec}, S_{prim} \mid I, S_f) dS_{sec}$$

with respect to the posterior probability, $P(S_{sec}, S_{prim} \mid I)$. In practice, the posterior probability $P(S_{sec}, S_{prim} \mid I, S_f$ is factored into two terms (and a constant denominator) using Bayes theorem: the likelihood, $P(I \mid S_{sec}, S_{prim})$, which is determined by the generative model for the image measurements (see figure 2.1), and the prior probability, $P(S_{prim})$. A simple Bayesian maxim summarizes the above calculation: Condition the joint probability on what we know, and marginalize over what we don't care about[2]. As seen in the color constancy section below, we have prior knowledge of the illuminant spectrum, we measure the shape and image color, we don't care about the illumination direction, and we want to estimate the surface color properties that minimize color judgment errors.

Pattern inference theory, ideal observers, and human vision

In order for the pattern inference theory approach to be useful, we need to be able to construct predictive theories of visual function which are amenable to experimental testing.

How do we formulate and test theories of vision at the functional level within a Bayesian pattern inference theory framework? Tests of human perception can be

1. The cost or *risk* $R(\Sigma; I)$ of guessing Σ when the image measurement is I is defined as the expected *loss*:

$$R(\Sigma; I) = \int_S L(\Sigma, S) P(S \mid I) dS$$

with respect to the posterior probability, $P(S \mid I)$. The best interpretation of the image can then be made by finding the Σ which minimizes the risk function. The loss function $L(\Sigma, S)$ specifies the cost of guessing Σ when the scene variable is S. One possible loss function is $-\delta(\Sigma - S)$. In this case the risk becomes $R(\Sigma; I) = -P(\Sigma \mid I)$, and then the best strategy is to pick the most likely interpretation. This is standard *maximum a posteriori estimation* (MAP). A second kind of loss function assumes that costs are constant over all guesses of a variable. This is equivalent to marginalization of the posterior with respect to that variable.

2. This Bayesian maxim is due to James Coughlan

based on hypotheses regarding constraints contained in: the two components of the generative model, 1) the prior $p(S)$ and 2) the likelihood $p(I|S)$; 3) the image model $p(I)$; 4) the posterior $p(S|I)$; or 5) the loss function. The first four can be called information constraints, whereas the fifth can be called a decision constraint. These levels can be translated into tests of: 1) prior knowledge and assumptions of scene structure; 2) model of the observer's image formation process; 3) what image measurements and image coding does the observer do; 4) the information about a scene available to the observer given an image; 5) the decisions and strategies used by the observer.

It is at the level of the posterior that provides the most complete quantitative model of the information available to an observer to perform a task. Whether this information is used optimally or not depends on whether the observer's loss function is matched to the particular task. Thus it is important to investigate more than one task that uses the same posterior in order to attribute a loss of information to the observer's posterior rather than a loss due to an inappropriate loss function. In all cases, we want to design our experiments to focus on the constraint hypotheses. For example, if we are interested in testing hypotheses about the observer's prior distribution on light source direction, then the experimenter could focus on simple ambiguous scenes whose interpretation depends on the light source direction. In these scenes the prior is expected to dominate.

However, the fact that testing at the level of information constraints uses information common to many decisions has a practical side-effect: we can define subdomains of S and I that are more easily implemented in the laboratory, and yet use the same posterior. This allows the experimenter to focus on testing the interesting predictions of the hypotheses.

As we've noted above above, in psychophysical experiments, one can: a) test at the functional or constraint level–what information does human vision avail itself of?, or; b) test at the mechanism level–what neural subsystem can account for performance? Because of its emphasis on modeling natural pattern representations and transformations, Pattern Inference theory is of primary relevance to hypotheses testable at the former level, (e.g. hypotheses about the representations appropriate for inference). Signal detection theory is of primary importance for the latter as we will see in the motion example below, where SDT provides the tools for rigorous tests of neural representation. In both cases, using an ideal observer allows a simple and meaningful comparison to be made between human and ideal performance.

The primary use of an ideal observer in an experimental setting is to provide a measure of the information available to perform a task that can be used to normalize the performance of any other human of model observer. Reporting human and model performance relative to ideal performance allows a straight forward comparison of the results from completely different tasks and visual cues. In other words, it allows comparisons like "Is the visual system better at

processing edges or shading information?" [42]. A standard way of performing this normalization is to compute efficiency, which is a measure of the effective number of samples used by the observers [11, 2].

This normalization function also provides a criterion by which to judge whether two models are functionally equivalent, and when a model can be firmly rejected. If two models produce the same efficiency curves for all relevant changes in task and stimuli, then the two models are equivalent. Notice, however, that two models can be equivalent by this criterion and yet could employ very different computations at the level of mechanism. This also allows us to construct models that are functionally equivalent to the human visual system. Everywhere human performance deviates from ideal, we modify the ideal observer to discard equivalent information and no more. Notice that this construction has a test of the viability of a model for human performance built into it. Anytime a human observer can outperform a model observer on a task, that model can be eliminated as a possible model for the human visual system.

In the next section we will look at an in-depth example of an application of the ideal observer approach to the problem of the determining the mechanisms used by the visual system to detect motion.

Ideal observer analysis: Mechanisms of motion measurement

Experimental studies of human perceptual behavior are often left with a crucial, but unanswered question: To what extent is the measured performance limited by the information in the task rather than by the perceptual system itself? Answers to this question are critical for understanding the relationship between perceptual behavior and its underlying biological mechanisms. Signal detection theory provided an answer through ideal observer analysis.

To show the power and limitations of this approach consider an example of a recent application (by one of the authors) of classical signal detection theory to the problem of visual motion measurement.

A model for motion detection

When a person moves relative to the environment, the visual image projected onto the retina changes accordingly. Within small regions of the retina and for short durations this image change may be approximated as a two-dimensional translation. The set of such translations across the visual field is termed the Optic Flow field. Having described a simple approximation of the otherwise complicated time-varying retinal image, the question remains whether the human visual

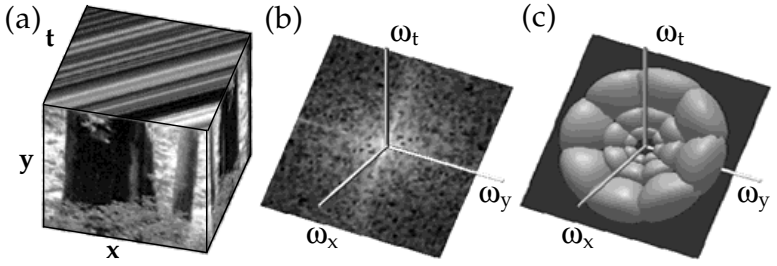

Figure 2.2: A translational motion detector. **a**, Space-time luminance pattern of an image translating to the right. This is a representation of the intensity information in the retinal image (the x- y plane) over time (t). The rightward motion can be inferred from the oriented pattern on the x-t face. **b**, The Fourier amplitude spectrum of the luminance pattern, represented by the intensity of points in a three-dimensional spatio-temporal frequency domain. Non-zero Fourier amplitudes are constrained to lie on a plane through the origin. The orientation of this plane uniquely specify the direction and speed of translation. **c**, Construction of a translation detector [39], illustrated in the Fourier domain. Pairs of balls symmetric about the origin indicate the Fourier amplitude spectra of band-pass filters whose peak frequencies lie in the plane. A translation detector can be constructed by summing the squared outputs of such filters.

system uses such an approximation. A number of physiological and psychophysical experiments have established that the mammalian visual system does contain mechanisms sensitive such local image translations [31], but these studies did not specify how local image translations might be measured by the visual system. One approach, initially due to Heeger [19] and later refined [18, 41] , uses fundamental properties of translating signals to derive an estimator for the velocity of the translation. Consider an image sequence $I(x, y, t)$. If in a window of space and time $W(\vec{x}, t)$ (e.g. gaussian) the image motion can be described as a translation, then $I(x, y, t) \approx \sum_{ij} w_{ij}(\vec{x} - \vec{x}_i, t - t_j)I(\vec{x} - \vec{v}_{ij}t, t)$. It is easy to show that the spatio-temporal (3-D) Fourier transform of one windowed region is given by

$$\mathcal{F}\{w_{ij}(\vec{x}-\vec{x}_i, t-t_j)I(\vec{x}-\vec{v}_{ij}t, t)\} = S(\vec{\omega_x}, \omega_t) = W(\vec{\omega_x}, \omega_t) \otimes (S_I(\vec{\omega_x})\delta([\vec{\omega_x} \;\; \omega_t]^T[\vec{v}_{ij} \;\; 1]))$$

Note that the delta function term is an equation for a plane in the Fourier domain specified by $[\vec{\omega_x} \;\; \omega_t]^T[\vec{v}_{ij} \;\; 1] = 0$. Thus the equation says that local image translations in the Fourier domain are characterized by the spatial spectrum of the image projected onto a plane whose orientation is uniquely specified by the velocity of the translation, which is convolved by the Fourier transform of the windowing function. For a Gaussian windowing function, the result is easy to state: translations are specified by blurred planes (or pancakes) in the Fourier domain. Figure 2.2a and b illustrate this without the windowing function. Given this description, a simple velocity detector can be constructed by pooling the outputs of spatio-temporal filters whose peak frequencies lie on a common plane (e.g. see figure 2.2c). Because the phase spectrum is not required for the velocity

estimates, a noise resistant [39] detector can be built by pooling the outputs of filters that compute the *energy* within a region of spatio-temporal frequency (e.g. like complex cells in V1). For a windowed signal S, the output R of the detector is given by

$$R = \sum_j \sum_{\vec{\omega_x}, \omega_t} a_j |F_j(\vec{\omega_x}, \omega_t)|^2 |S(\vec{\omega_x}, \omega_t)|^2$$

where F_j denotes whose peak frequency lies on the plane specified by the signal. Within this simple theory, we have a choice of the weights a_j. Given a particular image, only some of the filter bands F_j will contain the signal, and responses from filters not containing the signal will be solely due to noise. Thus a "smart" detector can improve detection performance by adjusting its pooling weights to match the signal. On the other hand, a good non-adaptive detector can be built by optimizing the weights for the expected (average) signal. This leads to a detector that pools over all spatial frequency orientations in a plane, because the expected spatial signal is approximately isotropic. We wanted to test whether an adaptive or fixed pooling power detector is a good model of human motion detection.

Notice the putative motion detection mechanisms have not been motivated from within signal detection theory. Rather they were motivated via a simple approximation and signal processing issues.

Testing the model

Finding a task for which the model is ideal

Signal detection theory can be used to assess the feasibility of such a model. To do so, we use the fact that the model we are interested in is an ideal observer for *some* task and stimuli. The idea is that if we have human observers perform the optimal task for the model, then if the model is a good description: 1) Humans should be good at the task 2) the model should predict errors on related tasks. Schrater et. al. [35] have recently shown that the putative motion detectors are ideal observers for detecting a class of novel stochastic signals added to Gaussian white noise. The stochastic signals are produced by passing Gaussian white noise through the filters used to construct the motion detector. In general, a detector which computes the Fourier energy within a filter is an ideal observer for stochastic signals generated by passing Gaussian white noise through the filter. Thus, by varying the number and placement of filters, we can produce motion stimuli that are consistent with a single translational velocity and have various spatial frequency spectra, or stimuli that are consistent with multiple velocities. Examples of some filters and stimuli are shown in figure 2.3.

So how do we go about testing our model by detecting these stochastic stimuli?

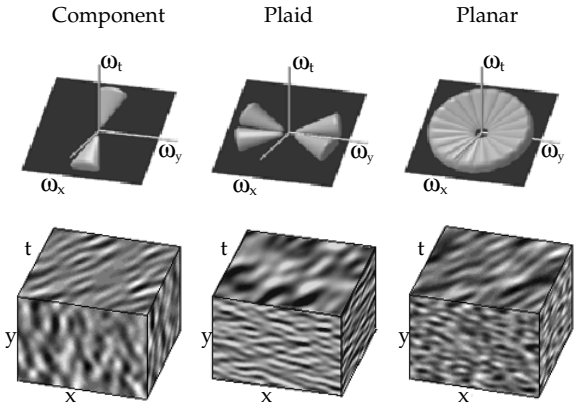

Figure 2.3: Filter sets and examples of their corresponding signals. The top row depicts level surfaces (65% of peak response) of the three different filter sets used to generate stimuli. The bottom row depicts space-time luminance patterns of signals produced by passing spatio-temporal Gaussian white noise through the corresponding filter sets. **(a)** The "component" stimulus, constructed from a spatially and temporally band-pass filter. The x-y face of the stimulus shows structures that are spatially band-pass and oriented along the x axis. The orientation of structures on the x-t face indicates rightward motion. **(b)** The "plaid" stimulus, constructed from two "component" filters lying in a common plane. The x-y face of the stimulus shows a mixture spatial of spatial structures with dominant orientations close to the y axis. **(c)** The "planar" stimulus, constructed from a set of 10 "component" filters lying in a common plane. The stimulus is spatially band-pass and isotropic (x-y face), and exhibits rightward motion (x-t face).

The ideal observer analysis gives immediate simple testable predictions: 1) Under the adaptive model as we vary the spatial structure of the motion stimulus, human performance should be constant relative to the ideal for each stimulus. 2) Under the fixed model, we should see predictable variations in performance. 3) If the model is false, we should be terrible at the task.

Testing the predictions

Translating these predictions into the detection task, the adaptive model predicts that any configuration of Fourier energy on the plane should be equally detectable. To test this prediction, we had observers vary the total energy of one of the three signals shown in figure 2.3 added to white noise until just detectable. The thresholds are plotted as signal power to noise power ratios in figure 2.4a. Note that the thresholds are lowest for detecting the "planar" signal with energy spread equally across a plane, followed by "plaid", with two bands, followed by the "component" filter, with only one band. However, unlike the non-stochastic stimuli used in previous signal detection experiments, here the signal to noise energy measure does

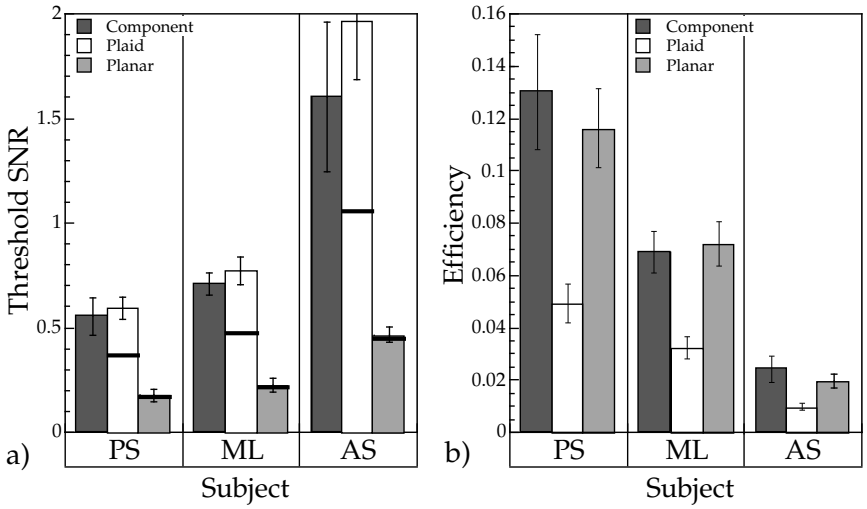

Figure 2.4: (a) Detection performance of three subjects for the three stochastic signals of Fig. 2.3. Threshold signal to noise ratio (SNR) for 81.1% detectability. SNR is calculated as the ratio of the signal power to the background noise power. Heavy black lines indicate predictions for ideal summation, derived from the component condition thresholds. **(b)** Detection efficiencies for the three stimulus types. Efficiencies are plotted in proportions, with 1.0 reflecting perfect performance; that is, performance matching that of an ideal observer tuned to the structure of the signal in the stimulus (different for each stimulus type). The differences between the efficiencies of the pattern stimuli (plaid and planar stimuli) and the component stimulus provide a quantitative measure of summation of the pattern's components.

not capture the subject's relative performance on the three stimuli. Looking at the threshold data, we might be tempted to conclude that planar stimuli are most easily detected and "component" and "plaid" stimuli are comparable in detectability. In fact, if we correctly normalize the observer's thresholds by the ideal observers thresholds for each stimulus to compute an efficiency measure, then we find that the "planar" and "component" stimuli are about equally detectable, whereas the "plaid" stimulus is much less detectable. Efficiencies are plotted in figure 2.4b.

The results suggest that the human visual system has band-pass filters similar to the "component" stimulus filter (and similar to V1 complex cells), and similar to the "planar" stimulus filter, but not to partial tilings of the plane. Thus the predictions of the fixed detector model have been confirmed, while the predictions of the adaptive detector have been contradicted. Notice also that the information normalization provided by the ideal observers is not superfluous. An attempt to compare detection performance across stimuli in terms of common stimulus measures (e.g. Fourier energy or power spectral height) for detectability would lead to erroneous conclusions. In addition, note that the efficiencies on these stimuli are about 10%, which is not extremely high (e.g. $> 50\%$ has been found for some other detection tasks [7]) but do not rule out the model either. It is likely,

however, that the model does not capture all the important elements of motion detectors in the visual system (e.g. opponency).

From ideal observer to human observer model

The question remains how much of human motion detection can be accounted for by the simple fixed detector that pools over an entire plane. To address this question, we can turn from looking at the ideal observer for each stimulus and instead try to predict the inefficiencies in human performance using the fixed detector model.

To do this, we compared the detectability predicted by the fixed model on a set of five stochastic stimuli, three new and the "planar" and "plaid" stimuli above. The new stimuli were created by passing spatio-temporal white noise through three configurations of filters, illustrated in Fig. 2.5. The first is a plaid signal, similar to the plaid used above. The second is a "planar triplet", created by adding a component band to the plaid, in the same plane as the plaid, and the third is a non-planar triplet, created by adding a component band out of the plane of the plaid. Detection thresholds were measured using the same method as above. The model predicts improved summation for the planar triplet, relative to the plaid, but no improved summation for the non-planar triplet. We computed predictions for the detection thresholds of each of the stimuli by implementing a specific fixed power dectector model. The detector optimally summed energy over the band of frequencies contained in the planar stimulus from experiment 1. We assumed that the output of this detector was corrupted by the internal noise levels estimated from subjects' detection thresholds for the component stimulus in experiment 1. Figure 2.5 shows observer's thresholds compared to the model predictions for the five pattern stimuli used. Given the assumptions built into the model concerning the exact spatio-temporal frequency band covered by the planar power detector, the match is surprisingly good. That is, not only do the qualitative results follow the predictions of the planar power detector model, but the quantitative results are well fit by a pre-defined instantiation of the model (without fitting the parameters of the model to the data).

Although SDT worked well in the analysis of mechanisms of motion detection, we need a theoretical framework for which the signals can be any properties of the world useful for the visual behavior; for example, estimates of object shape and surface motion are crucial for actions such as recognition and navigation, but they are not simple functions of light intensity. Natural images are high-dimensional functions of useful signals, and arriving at decoding functions relating image measurements to useful signals is a major theoretical challenge. However, both of these problems are expressible in terms of Pattern theory.

So, in the next section, we focus on the first problem: How can we model the

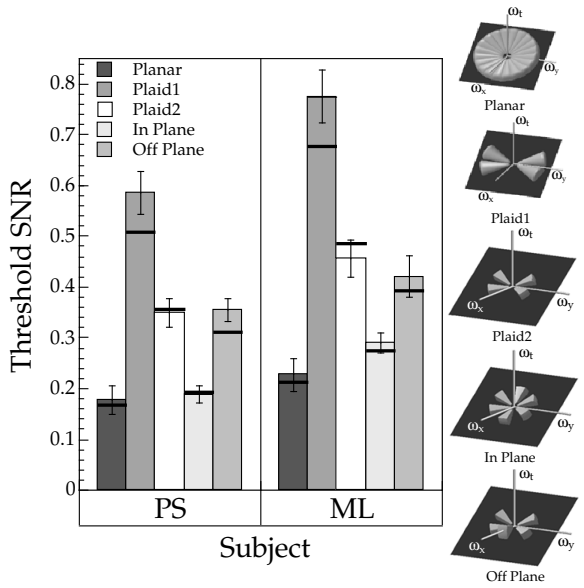

Figure 2.5: Threshold SNRs for detecting the five types of pattern stimuli replotted from experiments 1& 2, where Plaid1 in the legend denotes the plaid from the first experiment and Plaid2 from the second. Plaid1 differs from Plaid2 in that its energy is more diffusely spread over frequency. Black bands indicate the predictions of a planar filter, based on subjects' detection thresholds for the component stimulus used in experiment 1.

computations that have to be solved? This modeling problem can be broken down into synthesis: a) modeling the structure of pattern information in natural images; and analysis, b) modeling the task and extracting useful pattern structures.

Bayesian models: Color constancy

Robust object recognition relies on the estimation of object properties that are approximately invariant with respect to secondary variables such as illumination and viewpoint. Material color is one such property, and the phenomenon of perceptual color constancy is well-established (cf. [6] for a Bayesian analysis). For practical and scientific reasons, most laboratory studies of human color constancy have been limited to simple flat surfaces, or to the lightness dimension (see [1] and [5]).

Extracting color invariants from real surfaces is a more theoretically complex task, the wavelength information received by the eye being a function of the surface shape, material properties, and the illumination. Adding to this complexity is the fact that wavelength information also depends on reflected light from nearby

surfaces. Until recently, it was not at all clear whether human vision makes use of knowledge of 3D surface structure to infer surface color properties, which as we will see, involves a rather subtle computation. A recent study by Marina Bloj and colleagues has shown that human vision can indeed take into account 3D shape when inferring surface color [4]. In addition, their ideal observer analysis of the physical information available provided a quantitative account of human color matches.

Figure 2.6 illustrates the basic finding. A card consisting of a white and red half is folded such that the sides face each other. If the card's shape is seen as it truly is (a corner), the white side is seen as a white card, slightly tinted pink from the reflected light. However, if the shape of the card is made to *appear* as though the sides face away from each other (convex or "roof" condition), the white card appears magenta–i.e. more saturated towards the red[3]. Bloj et al. made quantitative measurements in which observers picked a comparison surface whose color best matched the white side of the target (see figure 2.7).

Notice that this is a simplified version of the natural task of determining the reflectance and illumination of an object in the presence of other objects. The problem is interreflections. The interreflections can be modeled pairwise. Then the color is determined by the illuminant, the reflectance functions of the surfaces, and the configuration of surfaces (geometry). Let us look at how to model the physics given a pair of surfaces.

Look at the illustration of two surfaces in figure 2.6. When two surfaces are concave with respect to a light source, the interreflections off of both surfaces provide a secondary source of illumination with different characteristics for both surfaces. As the angle between these surfaces decreases, the amount of inter-reflected light re-reflected off the other surface increases, and hence the spectrum of the reflected light off both surfaces changes. On the other hand, as we increase the inter-surface angle, the amount of inter-reflected light decreases until it reaches zero at 90 degrees. Because of the perspective ambiguity in interpreting a folded card as convex or concave, there are two interesting subcases of this continuum, one where the angle is acute and one where the angle is obtuse. These two cases are experimentally the most interesting because they yield completely different shape and reflectance attributions to the observation that one of the two surfaces was pink.

Let's see how to model the information for optimal inference. The primary variable of interest is the reflectivity ($S_{prim*} = \rho$) (measured in units of chroma).

3. The apparent switch in shape from corner to roof can be accomplished either by using a pseudoscope, a pair of dove prisms which effectively reverse the stereo disparities, or by using a monocular cue, such as linear perspective (see http://vision.psych.umn.edu/www/kersten-lab/kersten-lab.html).

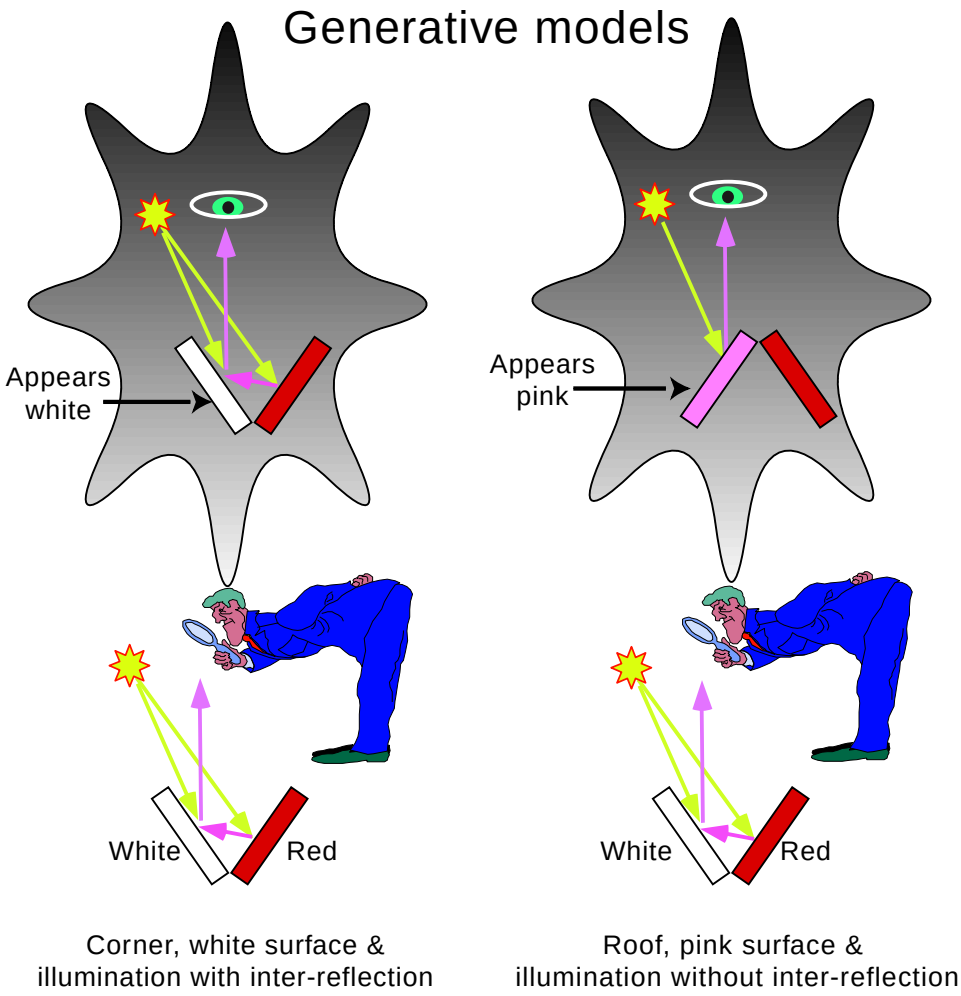

Figure 2.6: The white side of the folded card appears to either be pink or magenta, depending on the assumed shape of the card, i.e. whether the card is concave like a corner, or convex like a roof. The illumination model for a corner involves direct as well as inter-reflections, whereas the illumination model for the roof interpretation involves only direct lighting.

The likelihood is determined by either the one-bounce (corner) or zero-bounce model (roof condition) of mutual illumination. They assume that the shape is fixed by stereoscopic measurements, i.e. condition on shape ($S_f = roof\ or\ corner$). The one-bounce model yields the intensity equation for surface 1 (the "white surface"):

$$I_1(\lambda, x, \rho, E, \alpha_1, \alpha_2) = E(\lambda)\rho_1 * (\lambda)[cos\alpha_1 + f_{21}\rho_2(\lambda)cos\alpha_2]$$

where the first term represents the direct illumination with respect to the surface and the second term represents indirect (mutual) illumination due to light reflected from the red side (surface 2) [8]. $f_{21}(x)$ is the form factor describing the extent to which surface 2 reflects light onto surface 1 at distance x from the vertex [13]. The angles α_1 and α_2 denote the angle between the surface normal and the light source direction for surfaces 1 and 2 respectively.

For the zero-bounce generative model (roof condition):

$$I_1(\lambda, x, \rho, E, \alpha) = E(\lambda)\rho_1 * (\lambda) cos\alpha_1$$

These generative models provide the likelihood function. Observers do not sense I_1, but have a measure of color called chroma, modeled by the capture of the light by the retinal cones followed by a transformation, which we will denote as a function $C(\alpha, x, \rho, E) = f(\vec{K}(\lambda) \cdot I_i(\lambda, x, \rho, E, \alpha))$, where $\vec{K}$ denotes the three cone action spectra, and $\cdot$ denotes inner product. Thus the likelihood of observation C_{obs} being due to a color patch $C^i(\alpha, x, \rho^i, E)$ in the presence of some additive measurement noise is given by: $p(C_{obs}|\rho^i, x, \alpha, E) = K \exp(\frac{-0.5*(C_{obs} - f(\vec{K} \cdot I_i))^2}{\sigma^2})$. Now marginalize this conditional probability with respect to illumination direction and space (i.e. $S_{sec} = \{\alpha, x\}$) assuming uniform priors, and assume a priori (built-in) knowledge of the illuminant spectrum $E(\lambda)$ of daylight. Matching errors to the i_{th} patch are predicted by:

$$P(\rho_1^i|C_{obs}) \propto \sum_{\alpha} \sum_{x} exp - |C_{obs} - C^i(\alpha, x)|^2|/2\sigma^2$$

assuming a uniform prior on C_{obs} where σ is determined by the matching noise. (See [4] for details).

Experimental and theoretical matches are shown in figure 2.7. To a first approximation, the separation and spread of the observers' matches are predicted well by an observer which is ideal apart from an internal matching variability determined by σ. In other words, human matches are "ideal-like".

There are a number of important points to be made here. Note that the surfaces and lighting were carefully chosen to provide the bare minimal model for the natural visual task of inferring surface reflectance in the presence of interreflections. This reduction in complexity is important in that it allowed a highly controlled experimental test of the basic question. In addition, note that the one additional noise parameter that was used to model the observer's deviations from ideal is in fact not a free parameter (i.e. it was not fit to the data). Instead, the noise parameter was estimated from a separate color matching experiment.

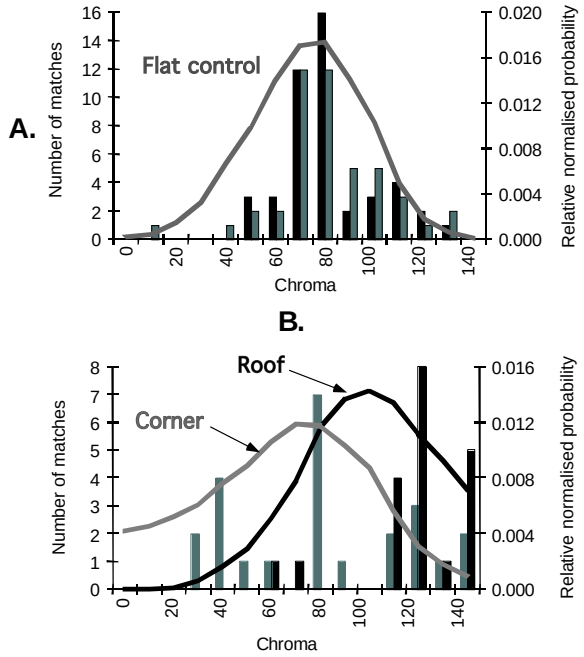

Figure 2.7: Panel A shows the distribution of human subject matches to a flat card, neither roof nor corner condition. Variability of matches is modeled as the standard deviation, σ of "matching noise". Panel A shows the distribution of matches under the two experimental conditions. Figure adapted from Bloj et al. Permission from MacMillan needed.

Summary and conclusions

We have discussed the problem of developing and testing quantitative models of human visual behavior. To that end we distinguished modeling function from modeling mechanism. The ideal observer plays a key role for both levels as:

- The information normalization tool for tests of visual mechanism.
- The default model for functional models of natural vision.

In testing models of vision, we emphasized the fundamental role of the ideal observer in interpreting human and model performance. The ideal observer provides a fundamental measure of the information available to perform a task, and thus serves to normalize human performance relative with respect to the task. Ideal observers can be used to define a task-independent measure of performance (efficiency), provide a measure of the functional equivalence of models, and serve as a default model to be modified by experiment. We described experiments of Schrater et al. supporting neural mechanisms specialized for the measurement lo-

cal image velocities, which are equivalent to specific sums of sets of complex cells in cortical area V1 [35].

By assuming that vision can be described as a process of decoding scene properties from images, we can use the approach of Pattern Inference Theory to develop ideal observers that serve as the starting point and comparator for a models of functional vision. A key point is that the importance of modeling the generative or forward process of natural image pattern formation (or encoding). The results of Bloj et al. [4] showed that the visual system takes into account knowledge of inter-reflected light in determining surface color. ¿From this perspective, theories of human visual performance can be developed iteratively from the ideal observer down (cf. [29]). This may be a more tractable strategy than to build models of system function bottom-up from mechanism components.

References

1. Adelson, E. H. (1999). Lightness Perception and Lightness Illusions. In M. Gazzaniga, M. S. (Ed.), *The New Cognitive Neurosciences*(pp. 339-351). Cambridge, MA: MIT Press.
reduction of
2. Barlow, H. B. "A method of determining the overall quantum efficiency of visual discriminations.". *J. Physiol. (Lond.)*. 160, 155- 168. 1962.
3. Bishop, C. M. (1995). **Neural Networks for Pattern Recognition**. Oxford: Oxford Univeristy Press.
4. Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception via mutual illumination. *Nature*, **402**, 877-879.
5. Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences USA*, **96**, , 307-312.
6. Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *J Opt Soc Am A*, **14**, (7), 1393-411.
7. Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, **214**, 93-94.
8. Drew, M., & Funt, B. (1990). Calculating surface reflectance using a single-bounce model of mutual reflection. *Proceedings of the 3rd International Conference on Computer Vision* Osaka: 393-399.
9. Eagle, R. A., & Blake, A. (1995). Two-dimensional constraints on three-dimensional structure from motion tasks. *Vision Res*, **35**, (20), 2927-41.
10. Eckstein, M. P. (1998). The lower efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, **9**, 111-118.
11. Fisher, R. A. (1925). **Statistical Methods for Research Workers**, Edinburgh: Oliver and Boyd.
12. Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, **368**, 542-545.
13. Foley, J., van Dam, A., Feiner, S., & Hughes, J. (1990). **Computer Graphics Principles and Practice**, (2nd ed.). Reading, Massachusetts: Addison-Wesley

Publishing Company.

14. Geisler, W. "Sequential Ideal-Observer analysis of visual discriminations". *Psychological Review*. 96,(2), 267-314. 1989.
15. Green, D. M., & Swets, J. A. (1974). **Signal Detection Theory and Psychophysics**. Huntington, New York: Robert E. Krieger Publishing Company. 1974.
16. Grenander, U. (1993).**General Pattern theory**, Oxford Univ Press.
17. Grenander, U. (1996). **Elements of Pattern theory**. Baltimore: Johns Hopkins University Press.
18. Grzywacz, N. M. & Yuille, A. L. A model for the estimate of local image velocity by cells in the visual cortex. *Proc. Royal Society of London A*, **239**, 129–161, (1990).
19. Heeger, D. J. Model for the extraction of image flow. *J. Opt. Soc. Am. A*, **4**, 1455–1471, (1987).
20. Kersten, D. (1984). Spatial summation in visual noise. *Vision Research*, **24**, 1977-1990.
21. Kersten, D. (1999). High-level vision as statistical inference. In Gazzaniga, M. (Ed.), *The New Cognitive Neurosciences* Cambridge, MA: MIT Press.
22. Kersten, D., & Schrater, P. W. (2000). Pattern Inference Theory: A Probabilistic Approach to Vision. In Mausfeld, R., & Heyer, D. (Ed.), *Perception and the Physical World*(pp. Chichester: John Wiley & Sons, Ltd.)
23. Knill, D. C. (1998). Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision Res*, **38**, (11), 1683-711.
24. Knill, D. C. (1998). Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. *Vision Res*, **38**, (11), 1655-82.
25. Knill, D.C., and Richards, W. (Eds). (1996). **Perception as Bayesian Inference**. Cambridge University Press. .
26. Knill, D. C., & Kersten, D. K. (1991). Ideal Perceptual Observers for Computation, Psychophysics, and Neural Networks. In Watt, R. J. (Ed.), *Pattern Recognition by Man and Machine*(pp. 83-97). MacMillan Press.
27. Koch, C., & Segev, I. (1998). Methods in Neuronal Modeling : From Ions to Networks. Cambridge, MA: MIT Press, 671 pages.
28. Liu, Z., Knill, D. C., & Kersten, D. "Object Classification for Human and Ideal Observers". *Vision Research*. 35,(4), 549-568. 1995.
29. Liu, Z., & Kersten, D. (1998). 2D observers for human 3D object recognition? *Vision Res*, **38**, (15-16), 2507-19.
    Neural
30. Mumford, D. (1996). Pattern theory: A unifying perspective. In Knill, D. C., & W., R. (Ed.), *Perception as Bayesian Inference*(pp. Chapter 2). Cambridge:

Cambridge University Press.

31. Nakayama, K. Biological image motion processing: a review. *Vis. Res.*, **25**, 625–660, (1985).
32. Olshausen, B. A., & Field, D. J. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". *Nature*. 381, 607-609. 1996.
    Intelligent
33. Richards, W. E. (1988). **Natural Computation**. Cambridge, Massachusetts: MIT Press.
34. B. Ripley. "Pattern Recognition and Neural Networks". Cambridge University Press. 1996.
35. Schrater, P. R., Knill, D. C., & Simoncelli, E. P. (2000). Mechanisms of visual motion detection. *Nature Neuroscience*, **1**, 64 - 68.
36. Schrater, P. (1998). Local Motion Detection: Comparison of Human and Ideal Model Observers. Ph.D. thesis, Philadelphia: University of Pennsylvania.
37. Schrater, P. R., & Kersten, D. (1999). Statistical Structure and Task Dependence in Visual Cue Integration. Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling, Fort Collins, Colorado.
38. Simoncelli, E. P., Adelson, E. H., & Heeger, D. J. (1991). Probability Distributions of Optical Flow. Mauii, Hawaii: *IEEE Conf on Computer Vision and Pattern Recognition*.
39. Simoncelli, E. P. (1993). Distributed Analysis and Representation of Visual Motion. Ph.D., Cambridge, MA: Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science,
40. Simoncelli, E. P. (1997). Statistical Models for Images: Compression, Restoration and Synthesis. Pacific Grove, CA.: IEEE Signal Processing Society.
41. Simoncelli, E. P. & Heeger, D. A model of neuronal responses in visual area MT. *Vis. Res.*, **38**, 743–761, (1998).
42. Tjan, B., Braje, W., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, **35**, (21), 3053-3069.
43. Watson, A. B., Barlow, H. B., & Robson, J. G. (1983). What does the eye see best? *Nature*, **31,**, 419-422.
44. Weiss, Y., & Adelson, E. H. (1998). Slow and smooth: a Bayesian theory for the combination of local motion signals in human vision (A.I. Memo No. 1624). M.I.T.
45. Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D.C., K., & W., R. (Ed.), *Perception as Bayesian Inference*(pp. Cambridge, U.K.: Cambridge University Press.

46. Yuille, A. L., Coughlan, J. M., & Kersten, D. (1998). Computational Vision: Principles of Perceptual Inference. *http://vision.psych.umn.edu/www/kersten-lab/papers/yuicouker98.pdf*

47. Zhu, S.C., Wu, Y., and Mumford, D. (1997). "Minimax Entropy Principle and Its Application to Texture Modeling". Neural Computation. **9**(8).

Acknowledgement

Supported by NSF SBR-9631682 and NIH RO1 EY11507-001.