Network 3 (1992) 213–251. Printed in the UK

REVIEW ARTICLE

Could information theory provide an ecological theory of sensory processing?

Joseph J Atick†
School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA

Received 31 December 1991

Abstract. The sensory pathways of animals are well adapted to processing a special class of signals, namely stimuli from the animal's environment. An important fact about natural stimuli is that they are typically very redundant and hence the sampled representation of these signals formed by the array of sensory cells is inefficient. One could argue for some animals and pathways, as we do in this review, that efficiency of information representation in the nervous system has several evolutionary advantages. Consequently, one might expect that much of the processing in the early levels of these sensory pathways could be dedicated towards recoding incoming signals into a more efficient form. In this review, we explore the principle of efficiency of information representation as a design principle for sensory processing. We give a preliminary discussion on how this principle could be applied in general to predict neural processing and then discuss concretely some neural systems where it recently has been shown to be successful. In particular, we examine the fly's LMC coding strategy and the mammalian retinal coding in the spatial, temporal and chromatic domains.

1. Introduction

This review explores the use of information theory (Shannon and Weaver 1949) as a basis for a first principles approach to neural computing. The relevance of this theory to the nervous system ultimately derives from the fact that the nervous system possesses a multitude of subsystems that acquire, process and communicate information. This is especially true in the sensory pathways. One could use information theory to assess the efficiency of information representation in many of these pathways. This already has given some insight into computational strategies in simple neural systems (Bialek *et al* 1991a, Warland *et al* 1992). More interestingly one could argue, as we do in section 3, that efficiency of information representation in the nervous system potentially has evolutionary advantages (Attneave 1954, Barlow 1961, 1985, Uttley 1979, Srinivisan *et al* 1982, Linsker 1988, 1989a,b, Field 1987, Atick and Redlich 1990a,b, 1992a, Atick *et al* 1990, 1991, Bialek *et al* 1991b, see also Barlow 1989 and references therein) and that much of the processing in the early levels of sensory pathways might be geared towards building efficient representations of sensory stimuli in an animal's environment.

The above efficiency principle, formulated as an optimization problem, can be used as a *design* principle to predict neural processing. Starting with the natural

† Address after 1 July 1992: The Rockefeller University, 1230 York Ave, New York, NY 10021, USA.

0954-898X/92/020213+39$04.50

representation of environmental signals as sampled by the array of sensory cells, one can try to find the recodings needed to improve efficiency subject to identifiable biological hardware constraints. The several stages of processing required to cast incoming data into the optimal form can then be compared to the stages of neural processing observed in sensory pathways. This principle has been shown to successfully predict retinal processing in space–time and colour (Atick and Redlich 1990a,b, 1992a, Atick *et al* 1990, 1991), and there are encouraging signs that it could be equally successful in predicting some of the cortical computation strategies (Barlow 1989, Field 1989, Barlow and Foldiak 1989, Atick *et al* 1992). The approach just described can be termed 'ecological', since it attempts to predict neural processing from physical properties of the stimulus environment. Essential to the success of this programme is a quantitative knowledge of (statistical) properties of natural signals. Several studies on properties of natural stimuli are currently underway.

The organization of the review is as follows. We start in section 2 with a brief review of information theory cast in a language suited for our subsequent analysis. In section 3, we speculate on why efficiency of information representation could be an organizing principle underlying sensory processing. We then formulate this principle as an optimization problem and discuss how in general it might be solved. In sections 4 and 5 we analyse in detail some biological systems where information theory has been shown to predict the observed neural processing. In section four, we analyse the contrast-coding of the LMC cells in the blowfly compound eye (Laughlin 1981, 1989), while in section 5 we study the spatio-temporal (Atick and Redlich 1990a,b, 1992a) and colour (Atick *et al* 1990, 1991) coding of the mammalian retina. Our discussion on retinal processing is self-contained since in subsection 5.1 we have included a brief review of the relevant experimental facts on retinal coding in space, time and colour.

2. Information theory: a quick primer

Information theory evolved in the 1940s and 1950s in response to the need of electrical engineers to design practical communication devices. The theory, however, despite its practical origins, is a deep mathematical theory (Shannon and Weaver 1949) concerned with the more basic aspects of the 'communication process'. In fact, it is a framework for investigating fundamental issues such as efficiency of information representation and its limitations in reliable communication. The practical utility of this theory stems from its multitude of powerful theorems that are used to compute optimal efficiency bounds for any given communication process†. These ideal bounds serve as benchmarks to guide the design of better information systems.

In this section we give a brief review of information theory. This review is not intended to be a full account of the theory. It focuses primarily on one aspect of information theory, namely the effect of statistical regularities on efficiency of information representation. Other important aspects are ignored including the role of noise and the reliability of representation. However, this account is adequate to enable the reader with no prior knowledge of information theory to follow its subsequent applications to neural computing. Readers interested in further details are encouraged to consult the literature (Shannon and Weaver 1949, Gallager 1968).

† *Physicists might find these bounds reminiscent of the bounds set by the laws of thermodynamics on the* performance of heat engines.

2.1. Information sources and channels

In information theory any device, system or process that generates messages as its output is generically referred to as an *information source*. Although each source has its own representation that it uses to put out messages, generally speaking sources represent their messages as combinations of symbols selected from their *alphabets*, the list of all possible symbols they are capable of producing. The symbols are often called the *source symbols* or the *representation elements*. The choice of alphabet and the way the symbols are used to construct messages constitutes a representation or a code—source coding.

For example, a book in English can be thought of as the output of an information source—English language—whose alphabet is $A, \ldots, Z, +$ blank. Similarly, a neuron or a layer of neurons can act as an information source whose alphabet is the different neuronal response levels. Finally, an information source that is discussed often in this review is the visual environment, where the alphabet is the different grey levels of light pixels in the image mosaic. For simplicity, we introduce information theory for the discrete case, where there is a countable number, N, of symbols that can be produced by the information source†. In written English $N = 27$, while in an 8-bit grey scale imaging, $N = 2^8 = 256$.

An important fact about 'natural' information sources is that they never produce messages which are random combinations of their symbols. Instead, their messages tend to possess regularities or what is known as *statistical structure*. In other words, the way symbols are put together to form messages obeys certain statistical rules that are source specific. To begin with, information sources do not utilize their symbols with equal frequency. In long sequences of written English for example, E occurs at the rate of once in every ten letters while Z occurs only once in a thousand (Pratt 1942). In totally random sequences of English alphabet the frequency of occurrence would be once in every 27 for all the letters. The frequency of occurrence of source symbols is captured by the set of probabilities $\{P(m), m = 1, \ldots, N\}$.

More importantly, the selection of a symbol in a message is influenced by previous selections; i.e. symbols in a message are not statistically independent, instead there are intersymbol dependencies or correlations. Again, in English when a T occurs somewhere in a text it is very likely it will be followed by an H while it is very unlikely that it will be followed by a Q for example. This statistical influence can be quite significant and can extend up to many symbols. Mathematically, it is captured by conditional probabilities or equivalently by joint probabilities among symbols. For messages of length l symbols the joint probabilities are denoted by $\{P(m_1, \ldots, m_l)\}$ where m_i is the ith symbol within the message.

We model real information sources as stochastic systems (Papoulis 1984) that generate sequences of symbols subject to some statistical rules (see also Geman and Geman 1984, Kersten 1990). Since our knowledge of the statistical regularities of natural information sources is somewhat limited at this time, the rules we impose on our models represent only a subset of all regularities real information sources might possess. This is not necessarily a handicap since at any given stage in a sensory pathway, especially at the early stages, we suspect that *only* incomplete knowledge of statistical regularities of the stimulus source is available to neurons. For example, we shall see in section 5 that retinal cells receptive fields can be accounted for with knowledge of pairwise correlation function of input signals only. Thus an approximate

† This can always be achieved by an appropriate choice of discretization of source outputs.

model of natural scenes that generates luminosity pixels subject only to the constraint of a fixed pairwise correlation function may be sufficient for studying the retina. Of course, to predict the processing in the cortex, knowledge of more complex regularities is necessary.

Finally, another basic concept in this theory is the concept of an *information channel*, which is the medium through which messages from sources are transmitted or stored. Just like an information source a channel possesses a set of symbols, called *channel symbols*, which are used to carry the messages. The problem of mapping source symbols into channel symbols is referred to as the channel coding problem (Gallager 1968). For the sake of brevity in the present review we ignore all differences between source and channel coding and deal only with the generic problem of information representation regardless of where the coding is happening. This is justifiable especially since we are focusing on discrete noiseless information theory.

2.2. *Efficiency of information representation*

As mentioned earlier, one of the main concerns of noiseless information theory is quantifying efficiency of information representation. Intuitively, inefficiency can be attributed to the fact that information sources are constrained to obey statistical rules in constructing their messages. These rules build some degree of redundancy where, for example, many pieces in a message are *a priori* predictable from other pieces and from knowledge of the statistical structure. Also the presence of constraints implies that a source does not utilize its alphabet to its fullest capacity since the constraints limit the combinations of symbols that are allowed as output. Hence, a representation that possesses any statistical regularities is in many ways wasteful or inefficient. In this section we find a quantitative measure for this inefficiency.

To begin with, information theory attributes to each message in the ensemble M of all messages that can be produced by a source, a statistical quantity known as the *information* which is given by†

$$I(w) \equiv -\log_2 P(w) \tag{2.1}$$

where $P(w)$ is the probability of the message w normalized so that $\sum_{w=1}^{\mathcal{N}} P(w) = 1$, with $\mathcal{N}$ the total number of messages in the ensemble M. $I(w)$ is essentially a measure of 'surprise' or *a priori* 'unexpectedness' of a message. According to it, a message that occurs often $P(w) \sim 1$ has low surprise or information value $I(w) \sim 0$, while that which is unexpected has high information. This measure conforms to the usual editorial policy where rare events are given more attention than frequently occurring ones. However, we should emphasize that it ignores the semantic value of a message; in this theory, the unexpectedness of a message plays an important role but is distinct from the meaning of the message.

Averaging (2.1) over all messages in the ensemble M defines

$$H(M) = \sum_{w=1}^{\mathcal{N}} P(w)\ I(w) = -\sum_{w=1}^{\mathcal{N}} P(w) \log_2 P(w) \tag{2.2}$$

which is known as the *entropy* or average information per message. As is shown below, $H(M)$ is the mathematical object one needs to construct a quantitative measure of

† Since we use $\log_2$ the units of I are bits (or binary digits)/message.

efficiency. Its precise significance derives from the powerful theorems that were proven about it. For example, the source coding theorem (see e.g. Gallager 1968) shows that $H(M)$ is the *minimum* length in binary digits (bits) per source message that are needed on average to represent the outputs of the source. Immediately, this says that a representation is most efficient iff on average messages in the ensemble M are equal to $H(M)$ bits in length.

To see how $H(M)$ is used to define a quantitative measure of efficiency, we investigate its dependence on what one intuitively perceives as the cause of inefficiency, namely the statistical structure. For concreteness, we consider a representation where each message w is built out of a combination of l symbols, then $P(w) = P(m_1, \ldots, m_l)$. We examine the value of $H(M)$ as a function of the statistical structure of the source keeping the N symbols and the length l fixed. We show that $H(M)$ decreases the more statistical constraints the source has to obey in generating messages.

Consider first the case of a source that uses a representation where the symbols are statistically independent, i.e. the only statistical structure is that given by $\{P(m_i)\}$. In that case $P(m_1, \ldots, m_l) = P(m_1) \cdot P(m_2) \cdots P(m_l)$ and the entropy $H(M)$ can be written as a sum over the individual *symbol* (or *pixel*) entropies, $H(i)$,

$$H(M) = -\sum_{i=1}^{l} \sum_{m_i=1}^{N} P(m_i) \log_2 P(m_i) \equiv \sum_{i=1}^{l} H(i). \tag{2.3}$$

In general, however, the symbols are not statistically independent, so $P(m_1, \ldots, m_l)$ does not factorize into a product and the total entropy does not equal the sum of symbol entropies. Instead it satisfies†

$$H(M) \leqslant \sum_{i=1}^{l} H(i) \tag{2.4}$$

with equality iff the symbols are statistically independent. This means that statistical influence among symbols lowers $H(M)$ or the amount of information carried by those symbols, which is intuitive since in this case many of the symbols redundantly carry the same information.

The upper bound on $H(M)$ in (2.4) is not the absolute maximum since one can still look for the distribution $\{P(m_i)\}$ that maximizes the symbol entropy $H(i) = -\sum_{m_i=1}^{N} P(m_i) \log_2 P(m_i)$. Again, it is not hard to show that the maximum occurs when $\{P(m_i) = 1/N, \forall\, m_i\}$, or when the alphabets are utilized with equal frequency, as anticipated. The maximum this gives is

$$\max_{ss}(H(M)) = \max_{\{P(m_i)\}} \left(\sum_{i=1}^{l} H(i) \right) = l \log_2 N \equiv C \tag{2.5}$$

† To see how the proof goes consider the simple case of two symbols. Define the matrix $D_{ij} = P(m_i)P(m_j) - P(m_i, m_j)$, then using the fundamental inequality $x \geqslant \ln(1 + x)$ applied to $x = D_{ij}/P(m_i, m_j)$ we have the inequality $D_{ij}/P(m_i, m_j) \geqslant \ln(1 + D_{ij}/P(m_i, m_j))$. Multiplying this by $P(m_i, m_j)$ on both sides and summating on i and j remembering that $P(m_i) = \sum_j P(m_i, m_j)$ and $\sum_i P(m_i) = 1$ one arrives at $H(1) + H(2) \geqslant H(1,2)$. Generalizing this proof to arbitrary number of symbols is straightforward.

where the first maximization is over the full statistical structure, and the second is over the distribution $\{P(m_i)\}$. Thus, maximum entropy is achieved by a source that represents its messages such that no statistical regularities exist among the symbols. A representation with no statistical structure is one where the receiver's knowledge about what to expect is minimal and thus on average a message when received conveys maximum amount of 'surprise' or equivalently maximum amount of information $H(M)$.

The last equality in (2.5) defines another important information theoretic quantity, namely the *capacity* C of the representation or the channel, which is the absolute maximum information that l symbols selected from a list of N distinct alphabets could ever carry. Notice that $C = l \log_2 N = \log_2 N^l$, is the logarithm of the total number of messages, N^l, that the representation can carry. It can also be interpreted as the actual length of messages in binary digits. In English, $C/l = \log_2 27 = 4.73$ bits/letter, while the capacity of an 8-bit grey scale 256×256 pixel screen is $8 \times 256 \times 256$.

We are now ready to define a measure of efficiency: for any source with $H(M)$ using a representation of capacity C one useful measure of efficiency is

$$\mathcal{R} = 1 - H(M)/C \tag{2.6}$$

which is called the Shannon redundancy. Since $H(M) \leqslant C$, $0 \leqslant \mathcal{R} \leqslant 1$ with $\mathcal{R} = 0$ being the most efficient where $C = H(M)$. This measure has two interpretations. First, thinking of $H(M)$ as the *actual* amount of information transmitted and C as the *maximum* amount that could be transmitted, efficiency calls for using a channel where the transmitted rate $H(M)$ is as close as possible to the maximum rate C. Alternatively, since C is the average length of a message in bits and $H(M)$ is the smallest average length that can ever be achieved by any representation (source coding theorem), efficiency calls for finding a representation where the actual length C is as close as possible to minimum allowed $H(M)$.

In general, to improve efficiency one recodes the output of the source into a representation that uses C as close to $H(M)$ as possible. This data compression is achieved by discarding the structure that is *a priori* predictable from the messages (the statistical structure) leaving only the so-called 'textual' or non-predictable information. In principle, a coding strategy that takes advantage of all statistical regularities can compress the representation down to its minimal size, i.e. can allow the use of $C = H(M)$. In practice, it might prove computationally prohibitive to achieve the optimal compression. In general one tries to find a compromise between the complexity of the representation and its efficiency, for example by ignoring certain aspects of the statistical structure and concentrating on those regularities that are simple to disentangle and discard in recoding. Also in real information systems noise is always present. In that case it is not advantageous to eliminate the redundancy completely, since it is redundancy after all that distinguishes what is signal from what is noise. Information theory formulated for noisy channels can be used to find the best compromise. In our analysis of real neural coding in section five we use an effective approach to handle the noise without the need for developing the complicated machinery of information theory in the presence of noise. The more general approach for handling noise in early sensory processing can be found in (Atick and Redlich 1990a,b).

2.3. The cost of inefficiency

To illustrate the cost of inefficiency, it is helpful to start with an example. Consider the DNA of a fictional creature whose bases, A, T, C, G are assumed to occur with probabilities listed in the second column of table 1†. The problem is to find a coding that will store long sequences of this DNA on a computer disk economically. Since table 1 does not supply any knowledge of statistical structure beyond base probabilities, we have to treat this information source approximately as if no statistical influence among the bases existed, and deal with each symbol as an independent message. Then the entropy of this DNA is $H(M) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4}$ bits/base. This means that there exists a code that can represent this DNA's sequences with as few as $\frac{7}{4}$ bits/base. If we code the four bases into 00, 11, 01, 10, then the average length (or capacity) used is 2 bits/base which is greater than $H(M)$. However, if we code in the fashion illustrated in the third column of table 1, then the average length is $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4}$, which is exactly the entropy of the source and thus the most efficient code possible given base probabilities only.

Table 1. The probability distribution of the bases A, T, C, G of the DNA of a fictional creature and the two simple binary codes discussed in the text.

Symbol	$P(i)$	Code 1	Code 2
A	$\frac{1}{2}$	00	0
T	$\frac{1}{4}$	01	10
C	$\frac{1}{8}$	10	110
G	$\frac{1}{8}$	11	111

One might think that since the bases in code 2 are not of equal length that decoding sequences would be difficult. This is not true; the code by construction has a trivial decoding algorithm. In any sequence, a zero signals the end of a coded base; with one exception, where one does not encounter zero for three consecutive digits, in that case the base is G and the next digit is part of the next coded base. There is a general procedure for constructing these minimal redundancy codes known as Huffman coding which generalizes this trivial example to arbitrarily complicated real problems (see Gallager 1968).

Notice that code 2 is on average $\frac{1}{4}$ bits/base or 12.5% shorter than code 1. Thus if this creature has 10^9 bases in its DNA, code 1 effectively requires an additional $\frac{1}{4} \times 10^9 = 250$ Megabits or ~ 31 Megabytes to store the same information. Further savings in storage space could be achieved using a code that can discard other statistical regularities that this DNA might have, such as correlations among the bases. Of course, this would occur at the cost of increasing the complexity of the code.

The above example leads into the general question of the cost of inefficiency. In man-made systems, inefficiency usually means more storage space, more expenditure of transmission power, longer transmission times or in general larger bandwidths or dynamic range to transmit or store the same amount of information.

In biological systems the consequences of inefficiency are not as clear and they are most likely animal dependent. What one needs is a way to translate the information theoretic cost into a biologically significant cost to an animal. However, generally

† Never mind the fact that they violate Chargaff's rule.

speaking we suspect that most areas in the nervous system of many species could not afford to be inefficient, since invariably neurons have a limited response range (capacity or dynamic range) especially in comparison with the wide range of stimuli the animal encounters (Shapley and Enroth-Cugell 1984, Barlow *et al* 1987). Pooling its dynamic range resources, we believe the brain possesses a relatively limited number of states that it has to use to build representations of the great multitude of objects and events in its information-rich environment. Under such circumstances, an efficient representation can allow the brain to extract more information about its environment without the need to evolve to larger sizes.

In addition to savings in dynamic range, efficient representations could potentially facilitate certain cognitive tasks, such as associative learning (Barlow 1989) and pattern recognition. Actually, in higher animals we feel it is more likely that cognitive benefits are the driving force towards efficient representations. These issues are discussed in further detail in section 3.

2.4. *Types of inefficiencies*

There are two types of inefficiencies that one encounters in information systems. As we shall see shortly, both types can influence the computational strategies of real sensory neurons. Both were alluded to in our discussion above, here for future reference we exhibit them more explicitly. To do that, we rewrite the Shannon redundancy (2.6) as $\mathcal{R} = (1/C)(C - H(M))$ in the following equivalent form

$$\mathcal{R} = \frac{1}{C}\left(C - \sum_{i=1}^{l} H(i)\right) + \frac{1}{C}\left(\sum_{i=1}^{l} H(i) - H(M)\right) \tag{2.7}$$

where we have added and subtracted $(1/C)\sum_{i=1}^{l} H(i)$ to the definition of $\mathcal{R}$.

The two terms in the brackets in (2.7) explicitly quantify the contribution of the two forms of redundancy to $\mathcal{R}$. First if the alphabets are used with equal frequency then $\sum_{i=1}^{l} H(i) = l \times \log_2 N = C$ (2.5), and the first term in the bracket drops out. In general, however, $\sum_{i=1}^{l} H(i) < C$, and this term contributes positively to the redundancy. Second, if there are no intersymbol dependencies, then the total entropy $H(M)$ equals $\sum_{i=1}^{l} H(i)$ exactly and the second term in (2.7) vanishes. Typically, however, there are statistical relations among the symbols in which case $\sum_{i=1}^{l} H(i) > H(M)$ (2.4) and hence the second term contributes a positive amount to the redundancy. In a system where there is absolutely no redundancy $C = \sum_{i=1}^{l} H(i) = H(M)$ to make $\mathcal{R} = 0$.

To get a feel for the relative significance of the two types of inefficiencies, consider written English. There $C/l = \log_2 27 = 4.76$ bits/letter, while $H(i)$ computed using the well known probabilities of different symbols (Pratt 1942) is 4.03 bits/letter, which gives a redundancy of about only 15%. In general inefficiency due to unequal use of symbols is minor. The major source of redundancy comes from statistical correlations among symbols. For English, an estimate of $H(M)/l$ was first done by Shannon (1951) using a method that takes into account statistical correlations among the symbols. He found that the entropy is around 1.4 bits/letter. From $C/l = 4.76$, $H(M)/l = 1.4$ and $H(i) = 4.03$ we can see that redundancy due to intersymbol correlations in English is about 55%, making the total redundancy of written English

close to 70%†. The situation is very similar in many natural sensory information sources.

2.5. Minimum-redundancy vs minimum-entropy codes

The expression for $\mathcal{R}$ in (2.7) also makes explicit two classes of codes that we will refer to in our discussions on neural coding. A code that minimizes the full $\mathcal{R}$ is known as *minimum redundancy* code, while that which minimizes the part of $\mathcal{R}$ due to intersymbol correlations is known as *minimum entropy* code or *factorial* code‡. Minimum entropy codes minimize the difference $\sum_{i=1}^{l} H(i) - H(M)$. In the limit $\sum_{i=1}^{l} H(i) = H(M)$, they produce a representation where the symbols are statistically independent, so the probability of any message is given by the product of the probabilities of the symbols making up the message, i.e. joint probabilities factorize into products of individual probabilities (hence the name factorial code). If one insists on no loss of information, then factorial codes minimize $\sum_{i=1}^{l} H(i)$ subject to the constraint of fixed total entropy $H(M)$. We should emphasize that these codes are not by themselves redundancy reducing. In fact, from (2.7) we can see that these codes preserve the total redundancy by transforming redundancy due to correlations to redundancy due to unequal use of symbols.

The interest in minimum redundancy codes in engineering is clear; they allow the use of smaller dynamic range or smaller capacity. The reason factorial codes are also interesting is that usually after minimizing $\sum_{i=1}^{l} H(i)$ one can find trivial transformations to fit the coded messages into a channel with a smaller C, and thus they can be viewed as a convenient first step for achieving minimum redundancy codes. For a simple example of this type of two-stage coding applied to continuous signals see subsection 4.2.

In sensory pathways, we expect factorial codes to play an important role for two reasons: just as in engineering, factorial codes are excellent first steps towards redundancy reduction. This is especially true for natural stimuli where the most significant part of the redundancy is coming from intersymbol dependency. Second, factorial codes could have an intrinsic cognitive advantage beyond the fact that they enable the nervous system to use smaller dynamic range. Both issues are elaborated on in the next section.

3. Information theory as an ecological theory of sensory processing?

3.1. General remarks

The neural networks in the sensory pathways of animals are well adapted to processing signals from the 'natural' environment. One fact about these special stimuli which was discussed in subsection 2.1 is that they are never random; instead they tend to possess statistical regularities. For example, in natural images, due to the morphological consistency of objects, nearby pixels are very similar in their visual appearance. The luminosity profile in these images changes gradually in space and only abruptly at edges or borders. Similarly in time and colour there is continuity and smoothness.

† For estimates of redundancy in other western languages, see Barnard (1955).

‡ Elegant examples of factorial codes can be found in Barlow *et al* (1989) and Hentschel and Barlow (1991), see also Watanabe (1981, 1985).

This means that in natural images there is a high degree of spatio-temporal and chromatic correlation among pixels. Hence a pixel by pixel representation of natural scenes, which is the representation formed by the photoreceptor mosaic, is inefficient. This fact was well known to engineers in the television industry as far back as the fifties. In fact, the statistical studies on television signals that they conducted indicate that redundancy could run well in excess of 90% in natural images (Gouriet 1952, Kretzmer 1952, Harrison 1952, Schreiber 1956). The situation is expected to be similar for most other senses.

Given that natural stimuli come in a highly inefficient form, there are several reasons why the nervous system might invest some of its resources to recode incoming signals to improve efficiency. We present three potential benefits of efficient representations. The first is an advantage of strict redundancy reduction, while the other two are advantages of both redundancy reduced and minimum entropy representations†. These benefits, however, are not mutually exclusive and do not exhaust all potential advantages of efficiency. Furthermore the discussion in this section is somewhat heuristic; we hope to present a more mathematical analysis of the material in this section elsewhere.

3.1.1. Information bottleneck. It is possible that at some point along a sensory pathway there exists what may be termed an *information bottleneck.* This means that somewhere there exists a restriction on the rate of data flow into the higher levels of a pathway. This could arise from a limited bandwidth or dynamic range of a neural link, which is not unlikely given that neurons invariably possess limited response range (Shapley and Enroth-Cugell 1984, Barlow *et al* 1987). Alternatively, the limitation could be due to a computational bottleneck in the higher levels of the sensory pathway that restricts the number of bits of data per second that can be analysed in the object recognition process. An example of such limitation might be the 'attention bottleneck' which is suspected to occur somewhere between area $V4$ and the inferotemporal cortex IT (Van Essen *et al* 1991)‡.

Studies on the speed of visual perception (Sziklai 1956) and reading speeds (Kornhuber 1973), consistently give numbers around 40–50 bits/s for the perceptual capacity of the visual pathway in humans. This number can be interpreted as the maximum rate of visual information that can be processed by the deep layers of the visual pathway and is in a sense a measure of the bottleneck. On the other hand the rate at which visual data is collected by the photoreceptor mosaic is known to exceed 5×10^6 bits/s (Jacobson 1951). In order to fit the huge range of incoming signals into the limited capacity anticipated at higher levels a sensory pathway might have to perform a series of data compressions. One strategy for data compression in neural systems is redundancy reduction§ (Attneave 1954, Barlow 1961). Other strategies include noise filtering and generalization.

† At this stage we cannot tell which of the two strategies, redundancy reduction or minimum entropy, is more fundamental in the nervous system. However, since they are closely related we will continue to treat both on an equal footing under the banner of efficiency.

‡ Actually it is very unlikely that the bottleneck is abrupt. It is most likely happening through a gradual constriction of data flow.

§ Of course, if the animal's needs are very specific then it could develop specialized feature detectors—bug detectors—very early on in its pathways that are tuned for objects and patterns that are critical for its survival. Such detectors will cut down on the data rate since they discard almost everything they do not detect. In higher animals, where the needs are not very specific and where flexibility to changing environment is critical, a better strategy is one which recodes to improve efficiency without discarding a

3.1.2. Associative learning. Barlow (1989), argued that the way the nervous system represents objects and events in the environment might have dramatic implications to an animal's ability to perform associative learning. The idea is that for an animal to learn a new association between any two events, m_1 and m_2, the brain should have knowledge of the prior probability of occurrence or the *a priori* coincidence rate of m_1 and m_2. Without this information the animal cannot tell whether event m_1 has become a good predictor of m_2 or whether the joint occurrence of m_1 and m_2 (or m_1 followed by m_2) is consistent with the random coincidence rate, i.e. it cannot learn the association of m_1 and m_2†. What the animal needs is knowledge of the prior joint probability $P(m_1, m_2)$‡. Similar arguments apply for associations among any number of events.

However, knowledge of the prior probability of joint events in the environment is not easy to achieve. In general, there is a huge number of events and conjunctions among them. By any reasonable estimate, knowledge of the prior probabilities of all these conjunctions would require storage of an exponentially large set of numbers that far exceeds any estimate of brain storage resources. The only way out seems to be if the representation of events and objects in the brain is very special. In fact, if the representation is such that the elements are statistically independent—of course, until the association to be learned occurs—then the probability of any combination of them can be obtained very simply from individual probabilities, since in that case $P(m_1, \ldots, m_n) = P(m_1) \cdots P(m_n)$. Thus for any N events the N^n probabilities $\{P(m_1, \ldots, m_n)\}$ can be computed from knowledge of the N individual probabilities $\{P(m_i); m_i = 1, \ldots, N\}$§.

So the fact that the brain is finite in its resources suggests that a minimum entropy representation of the world might be necessary for it to perform a cognitive task essential for survival, namely associative learning.

3.1.3. Pattern recognition. The ultimate goal of any sensory pathway is pattern recognition: for its survival, an animal needs to acquire from its senses knowledge of the location and identity of all objects in its immediate environment. A third possible explanation for why a sensory pathway might choose to preprocess incoming signals to improve their efficiency is that efficiency might facilitate the pattern recognition process (see also Barlow 1985, Watanabe 1981, 1985).

Consider for instance the visual pathway. In the incoming representation, pixels are highly correlated and thus have low information value. A large number of pixels is needed to define any feature. An efficient representation, on the other hand, decomposes images in terms of elements that are statistically independent and thus

lot of information early on. In reality, a combination of the two mechanisms is in place. For example, an animal chooses a sensory sampling unit—acuity limit or resolution—below which it discards all data.

† In Pavlovian conditioning m_1 is the conditional stimulus while m_2 is the unconditional one.

‡ To be more precise, it needs knowledge of the conditional probability $P(m_2|m_1)$ which is related to the joint probability through $P(m_2|m_1) = P(m_1, m_2)/P(m_1)$. A high conditional probability $P(m_2|m_1)$ means that m_1 is a good predictor of m_2.

§ To take an example, imagine the situation where the visual pathway recodes images into a factorial representation. Then the probability of any scene can be computed easily from the product of probabilities of the individual elements that it activates. This scene probability can be thought of in two ways, one as the probability of some complex stimulus and two as the joint probability of the features that make up the stimulus. Thus factorial codes in vision provide the visual pathway with a simple way to compute joint probabilities of visual features.

necessarily more informative elements. These elements are the features or the 'vocabulary' from which natural images can be assembled most economically. It is possible that these building blocks, arrived at by pure statistical considerations, are closer to the patterns and objects an animal needs to recognize in its environment and hence a representation that uses them could simplify the subsequent pattern recognition process.

Independent of whether the visual system takes advantage of efficiency for pattern recognition, it is of interest to find what the features in efficient representations of natural images turn out to be. This is a concrete proposal, since starting with a database of natural images, one can look for transformations that drive $\mathcal{R}$ or some variant of it down. One promising approach for doing this is to use neural networks which can be trained using unsupervised learning algorithms that incrementally improve efficiency of representation as the network is exposed to more examples of natural images. Some unsupervised learning algorithms that achieve this in some simple settings have appeared in (Goodall 1960, Hinton and Sejnowski 1983, Pearlmutter and Hinton 1986, Barlow and Foldiak 1989, Redlich 1991, Atick and Redlich 1992b).

3.2. An optimization problem

In this section we formulate the principle of coding to improve efficiency as an optimization problem. For concreteness, we focus on visual processing. We make the hypothesis that the visual system is concerned with building a minimum entropy or factorial representation of the natural world†. What this means is that the visual system has to map the photoreceptor signals, which are highly correlated, to a representation where the elements are statistically independent. It is unlikely that any system could achieve this in one recoding. It is more likely that it would have to work in an iterative scheme that tries to improve efficiency by successively eliminating more complex forms of correlations. For instance, we shall see in subsections 5.2 and 5.3 that if at the first stage one insists on eliminating only second order statistics ignoring all the higher order regularities, one arrives at filters with properties that are close to those observed in the retina. It is then conceivable that the elimination of more complex statistical structures could lead to processing similar to that found in the primary visual cortex.

To begin with, let $\{L_i, i = 1, \ldots, n\}$ denote the activities of the n neurons in the input layer and $\{O_i, i = 1, \ldots, l\}$ the corresponding activities in the output layer. (l is not necessarily equal to n). The response of the output neurons is assumed to be some general function of the input activities:

$$O_i = K_i(L_1, \ldots, L_n) \qquad \forall i. \tag{3.1}$$

The input and output layer could be any two consecutive stages along the visual pathway. The question is then how should the recoding functions $\{K_i\}$ be chosen in order to achieve the desired statistical independence?

In subsection 3.2, we have seen that a recoding that minimizes the sum over pixel entropies $\sum_{i=1}^{l} H(O_i)$ to its absolute minimum while keeping the total entropy fixed achieves statistical independence. In general, one may not be able to find the $\{K_i\}$ that achieves the absolute minimum. For this reason, we define a fitness or

† Since by a simple transformation we can also achieve minimum redundancy, the results of this section are equally relevant to minimum redundancy coding.

energy functional, $E\{K_i\}$, that grades different recodings, $\{K_i\}$, according to how well they minimize the sum of pixel entropies without loss of information. A recoding is considered to yield an improved representation if it possesses a smaller value for E. The simplest energy functional for statistical independence is

$$E\{K_i\} = \sum_{i=1}^{l} H(O_i) - 2\rho[H(O_1, \ldots, O_l) - H(L_1, \ldots, L_n)] \quad (3.2)$$

where ρ is a parameter penalizing information loss. It can also be treated as a Lagrange multiplier in which case it enforces the constraint $H(O_1, \ldots, O_l) = H(L_1, \ldots, L_n)$ exactly. Any hardware constraint can be added to (3.2) with the appropriate Lagrange multiplier.

The optimal recoding can be found by solving the variational equations:

$$\frac{\delta E\{K_i\}}{\delta K_i} = 0. \quad (3.3)$$

In general, these equations are hard to solve if $\{K_i\}$ is allowed to be any arbitrary function. However, it is not clear that biology could implement recodings by arbitrary functions. A better approach would be to find the optimal solution for a restricted class of functions that are implementable by realistic layers of neurons. For example the retina to a good approximation performs a linear transform on the photoreceptor signals, so one could solve (3.3) for the class of linear functions.

An interesting simplification occurs when $\{K_i\}$ is restricted to the class of linear one to one ($l = n$) recodings, i.e. $O_i = \sum_{j=1}^{n} K_{ij} L_j$, $\forall i$. By a change of variables, keeping in mind that $P(O_1, \ldots, O_n)$ transforms as a density it is not hard to show that $H(O_1, \ldots, O_n) - H(L_1, \ldots, L_n) = \log \det \mathbf{K}$ independent of details of the statistical structure of natural scenes, where $\mathbf{K}$ stands for the matrix K_{ij}. The only knowledge of the statistics resides in the pixel entropies $\{H(O_i)\}$. In subsection 5.2, we solve (3.3) explicitly for this special class of codes. But first we discuss the statistics of natural scenes which are needed to compute $\sum_{i=1}^{l} H(O_i)$ in (3.2).

3.3. Statistics of natural scenes

Unfortunately only little is known at the quantitative level about the statistical properties of natural scenes. Some of that knowledge has come from the early work on the statistics of television images (Gouriet 1952, Kretzmer 1952, Harrison 1952, Schreiber 1956) and from the more recent measurements of the pairwise correlation function of natural scenes by Field (1987, 1989). Thus our model of natural scenes will have to be approximate.

The two-dimensional pairwise correlation function, or alternatively the spatial autocorrelator, is defined as

$$R(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle L(\boldsymbol{x}_1) L(\boldsymbol{x}_2) \rangle \quad (3.4)$$

where the brackets denote ensemble averaging over scenes or average over one large scene assuming ergodicity (Papoulis 1984). $L(\boldsymbol{x}_1)$, $L(\boldsymbol{x}_2)$ are the light levels above the mean level at two spatial points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. By homogeneity of natural scenes the autocorrelator is only a function of the relative distance, $\boldsymbol{X} \equiv \boldsymbol{x}_1 - \boldsymbol{x}_2$: $R(\boldsymbol{X})$.

One can thus define the *spatial power spectrum* which is the Fourier transform of the autocorrelator

$$R(f) = \int \mathrm{d}X \exp(\mathrm{i}f \cdot X) R(X). \tag{3.5}$$

For an ergodic system (Papoulis 1984), the power spectrum $R(f)$ is simply given by $L(f)L(-f)$, and therefore it is only necessary to take the Fourier transform of a scene $L(x)$ in order to compute the power spectrum.

This is what Field did, where he found that invariably for natural scenes

$$R(f) \sim 1/|f|^2 \tag{3.6}$$

which corresponds to a scale invariant autocorrelator: under a global rescaling of the spatial coordinates $x \to \alpha x$ the autocorrelator $R(\alpha x) \to R(x)$†. Although this scale invariant spatial power spectrum is by no means a complete characterization of natural scenes, it is the simplest regularity they possess.

The model of natural scenes that we adopt is one where the pixels $(L(x_1), \ldots, L(x_n))$ making up an image are chosen with a Gaussian probability distribution of the form

$$P(\mathbf{L}) = [(2\pi)^n \det(\mathbf{R})]^{-1/2} \exp\left[-\frac{1}{2}\mathbf{L} \cdot \mathbf{R}^{-1} \cdot \mathbf{L}\right]. \tag{3.7}$$

In writing this expression we have used upright bold-face symbols to denote matrices and vectors; R stands for the matrix $R_{ij} \equiv \langle L(x_i)L(x_j)\rangle$ and is given by the Fourier transform of (3.6), and L is the vector $(L(x_1), \ldots, L(x_n))$. The distribution in (3.7) is the one that gives maximum total entropy $H(\mathrm{L})$ consistent with the autocorrelator being R. In other words it is the distribution that incorporates no knowledge beyond what is specified by the autocorrelator, and hence is the one that most honestly reflects what we know about natural scenes. Equation (3.7) will be used in section 5.

4. Coding to improve efficiency: neural strategies

The number of examples of neural systems where a computational strategy to improve efficiency has been demonstrated, is growing (Laughlin 1981, Atick and Redlich 1990a, 1992a, Bialek 1990, Atick *et al* 1990). In this review, we only have space to discuss in detail two examples. These two illustrate coding strategies designed to deal with the two types of inefficiencies described in subsection 2.4. Our first example, discussed in this section, illustrates a coding scheme from the fly compound eye that eliminates inefficiency due to unequal use of neural response levels (Laughlin 1981). The second, to which we dedicate section 5, examines the mammalian retinal coding strategies in space–time and colour, which appear to be designed primarily to deal with inefficiency due to interpixel dependencies or correlations.

† To make the inverse Fourier transform of (3.6) well defined one has to use a low and high frequency cutoffs which physically correpond to 1/(size of the visual field) and 1/(resolution scale) respectively. These cutoffs violate the scale invariance of $R(x)$, which holds only as an approximate symmetry.

4.1. LMC gain control in the blowfly compound eye

The large monopolar cells (LMC) in the blowfly compound eye have been studied extensively over the last two decades (for reviews see Shaw 1984, Laughlin 1987). They are interneurons known to respond to contrast signals. These neurons, just like all other neurons, face a serious coding problem since they have a strictly limited dynamic range, i.e. they possess only a small number of distinguishable response levels. The question is how the LMC should choose its gain (or contrast sensitivity) so as to most efficiently represent the different contrast levels†.

If the LMC sets its sensitivity too high, inputs very often would saturate the response and much of the information about high contrast inputs would be lost. On the other hand, if the sensitivity is too low, the information about low contrast inputs would be lost. In both cases, the different output levels would be far from being equally utilized. In the first case, the higher output states are used much more often than the lower ones, while in the second case large parts of the output at the high end remain under-utilized. To achieve an efficient encoding, the LMC must choose its gain such that all response levels are used with equal frequency.

This problem was first analysed information theoretically by Laughlin (1981), here, we paraphrase his analysis. The first step in trying to discover the optimal code is to find out the statistical regularities of the input. In this case we only need to know the probability distribution of contrast signals occurring in the natural environment of the fly. Laughlin (1981) measured it from samples of horizontal scans of dry woodland and lakeside vegetation.

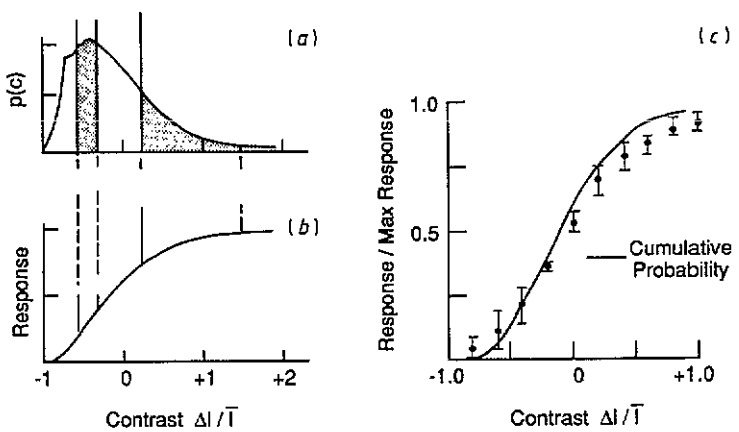

Figure 1. Probability distribution of contrasts, (*a*), in the fly environment from the measurements of Laughlin (1981). The contrast–response predicted by information theory is the cumulative probability map in (*b*). (*c*) is a comparison between the predicted response and that actually measured by Laughlin (1981) in the LMC.

Let us denote the input contrast signal by c, and use o to represent any one of the output or response levels, measured in some appropriate quantization units. The probability distribution for the input is $P(c)$ and it looks something like what is shown in figure 1(*a*), adapted from Laughlin (1981). The neural transfer function

† Here we are working at high luminosity so we can ignore the role of noise and treat the problem with the tools of noiseless information theory.

or the neural gain g defines a mapping from the input c to the output $o = g(c)$. To achieve optimal coding, the function g should be chosen such that the probability distribution of the output, $P(o)$, is constant for all output states o, i.e. $P(o) = \alpha$ for some constant α. Since the transform from the input c to the output o can be thought of as a change of variables, and since the probabilities transform as densities, then

$$P(o)\,\mathrm{d}o = P(c)\,\mathrm{d}c. \tag{4.1}$$

Setting $P(o) = \alpha$, we can integrate the resulting equation to find the transformation on the input needed to equalize the output probabilities†:

$$\begin{aligned} o = g(c) &= \frac{1}{\alpha}\int_{-1}^{c} \mathrm{d}c'\,P(c') \\ \frac{o}{o_{\max}} &= \int_{-1}^{c} \mathrm{d}c'\,P(c') \end{aligned} \tag{4.2}$$

which can be recognized as the cumulative probability map. Notice that the constant α is given by $1/o_{\max}$ since $o_{\max} = (1/\alpha)\int_{-1}^{1} \mathrm{d}c'\,P(c') = 1/\alpha$. Also, the sensitivity of the cell, defined as $\mathrm{d}o/\mathrm{d}c$, in this coding scheme is simply $P(c)$. So the neuron is most sensitive around the most probable input contrast, with its sensitivity dropping to zero as the signal c becomes improbable (see figure 1(*b*)).

Laughlin compared this predicted neural coding strategy to that found in experiments where he measured the response of the light-adapted LMC to sudden increments or decrements of light about the steady background level. The results of the comparison are shown in figure 1(*c*). The full curve is the cumulative probability computed from measured contrast probability distribution in the fly environment, while the dots with error bars are actual measurements of contrast response in the LMC. The dots represent the average of repeated responses to the same stimulus. The agreement is clearly very good.

4.2. Gain control in a layer of n neurons

In this section we generalize Laughlin's result to a layer of n neurons each receiving inputs from a spatial array of n sensory cells. If we denote the inputs from the sensory cells by $\{c_i, i = 1, \ldots, n\}$ and the response of the neurons by $\{o_i, i = 1, \ldots, n\}$, the question again is how to choose the gain function, defined by $o_i = g_i(c_1, \ldots, c_n)$, in order to use the neuronal output levels most efficiently. We will make the assumption that all the output neurons have the the same limited dynamic range.

The analogue of (4.1) here is

$$P(o_1, \ldots, o_n)\,\mathrm{d}o_1 \cdots \mathrm{d}o_n = P(c_1, \ldots, c_n)\,\mathrm{d}c_1 \cdots \mathrm{d}c_n. \tag{4.3}$$

In general, contrary to the one-neuron case, it is not obvious how to integrate (4.3) when we set $P(o_1, \ldots, o_n) = \alpha$. However, suppose that the neurons before choosing their gain function, coded the signals into a factorial representation, i.e. coded the

† The contrast signal is defined as $(I - I_0)/I_0$ where I is the intensity of a given pixel while I_0 is the average intensity within some visual window. This definition gives a contrast that cannot be smaller than -1.

input signals $\{c_i\}$ into $\{\gamma_i\}$ such that $P(\gamma_1, \ldots, \gamma_n) = P(\gamma_1) \cdots P(\gamma_n)$. Equation (4.3) can then be easily integrated to derive the necessary gain control on the resulting signals. The latter is simply given by the cumulative probability maps

$$\frac{o_i}{o_{\max}} = \int_{-1}^{\gamma_i} \mathrm{d}\gamma_i' P(\gamma_i') \tag{4.4}$$

for each neuron independently. Thus for this system, efficiency of output representation predicts the two stage coding shown in figure 2.

Figure 2. The problem is how to represent correlated signals most efficiently if the output neurons possess limited dynamic range. The simplest solution is shown as a two-stage process in this figure. In the first stage input signals are decorrelated to provide a minimum entropy code which can be followed by a simple cumulative probability map that functions as the gain control. The latter serves to fit the decorrelated signals into the limited dynamic range of each neuron independently.

To be more concrete we work out in detail the coding for the simple case of $n = 2$ and where the signals c_1 and c_2 are Gaussian signals with a correlator given by

$$\begin{pmatrix} \langle c_1\ c_1 \rangle & \langle c_1\ c_2 \rangle \\ \langle c_2\ c_1 \rangle & \langle c_2\ c_2 \rangle \end{pmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{4.5}$$

where r is a number < 1 characterizing the degree of overlap between the two channels, and the brackets denote ensemble averages. According to figure 2, the signals c_1, c_2 are first transformed to the decorrelated signals γ_+, γ_-:

$$\gamma_+ = \frac{1}{\sqrt{2}}(c_1 + c_2) \qquad \gamma_- = \frac{1}{\sqrt{2}}(c_1 - c_2). \tag{4.6}$$

The γ_+, γ_- signals are also Gaussian with variances $1 + r$ and $1 - r$ respectively. Thus the final transformation from γ_+, γ_- to o_+, o_- is given by a cumulative integral over a Gaussian which is simply related to the standard error functions. The net transformation for this system is

$$\begin{aligned} o_+ &\sim \operatorname{erf}\left(\frac{1}{\sqrt{2}} \frac{c_1 + c_2}{\sqrt{1 + r}}\right) + \text{constant} \\ o_- &\sim \operatorname{erf}\left(\frac{1}{\sqrt{2}} \frac{c_1 - c_2}{\sqrt{1 - r}}\right) + \text{constant}. \end{aligned} \tag{4.7}$$

Notice in the regime where the contrast signals are small in comparison with the square root of the variance, the response linearizes, $o_\pm \sim (c_1 \pm c_2)/\sqrt{1 \pm r}$, and the only effect of the gain control is to normalize the signals by dividing by the square root of the variance.

In the next section, we generalize the above minimum entropy code to the more realistic case of the array of retinal ganglion cells and will modify the coding to take into account the noise. However, we will ignore the gain control transform or the cumulative probability map and work purely within the linearized approximation.

5. Retinal coding strategies in space–time and colour

The mammalian retina is a rather unique neural system. It is a network that is complex enough so insight gained from understanding it promises to be useful in understanding other areas in the brain, yet it is still simple and isolated enough that quantitative experiments with clear outcomes can be performed. As such, the retina is ideal for developing and testing theoretical ideas on neural computations.

In this section, we start by giving a brief review of relevant experimental facts about the retina. We then explore the efficiency principle discussed in section 3 in the context of the retina. This gives a predictive framework that explains many of the experimental facts in subsection 5.1.

5.1. The retina: some relevant experimental facts

The retina is the thin tissue lining the back of the eyeball. As a neural network, it has feedforward architecture with three essential layers: photoreceptors, bipolar cells, and ganglion cells. However, it also has important lateral connections and interneurons acting within a given layer. The photoreceptor layer forms the input to this network, where photons from an image focused on the surface of the retina are captured and transduced into graded voltage signals. The output is built of spike trains generated by ganglion cells, and it propagates down the optic nerve to the LGN and subsequently to the visual cortex.

Since, here, we are only interested in functional properties of the retina, neither detailed connectivity of its network nor properties of cells other than photoreceptors and ganglion cells are of interest to us†. We simply think of the retina as a black-box processor whose input is the photoreceptors' activities and output is the ganglion cells' activities. This processor can be characterized by its *transfer function* which specifies how the output is related to the input, see figure 3.

The retinal transfer function is measured in single-cell recordings of ganglion cell outputs or inferred from psychophysical contrast sensitivity measurements subject to some plausible assumptions (see Shapley and Enroth-Cugell 1984 and references therein). In single-cell experiments, one finds that after adaptation to the light level the output of any given ganglion cell, measured as the rate of spikes in spikes/s, is to a good approximation given by a weighted sum of the photoreceptor activities over a small contiguous region on the surface of the retina known as the cell's *receptive field*, RF (figure 3). Thus the output of a ganglion cell whose RF is centred at $\boldsymbol{x}_i$ and at time t can be written as‡

$$O(\boldsymbol{x}_i, t) = \int \mathrm{d}\boldsymbol{x}' \, \mathrm{d}t' K(\boldsymbol{x}_i, \boldsymbol{x}'; t, t') L(\boldsymbol{x}', t') \equiv K \cdot L \tag{5.1}$$

where $L(\boldsymbol{x}', t')$ is the activity of the photoreceptor at location $\boldsymbol{x}'$ and at time t', while $K(\boldsymbol{x}_i, \boldsymbol{x}'; t, t')$ is the retinal kernel or retinal transfer function.

† For further information about retinal organization the reader should consult reviews on the subject (e.g. Davson 1980, Shapley and Enroth-Cugell 1984, Sterling 1990).

‡ The linear cells in cat are often referred to as the X cells, while in monkey they are known as the parvocellular cells which constitute about 80% of the ganglion cells in the retina. In monkey, they are considered to be part of a pathway that extends into the deep layers and is believed to be concerned with detailed form recognition (see e.g. Van Essen and Anderson 1988).

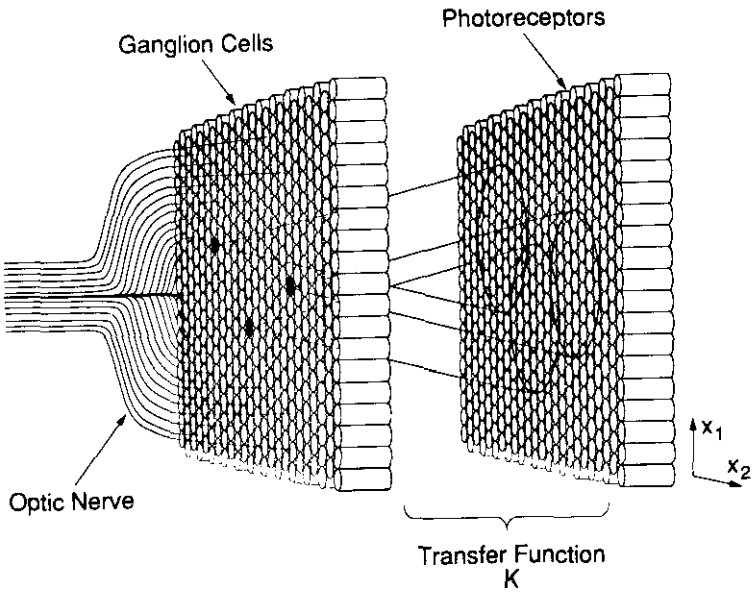

Figure 3. The retina as a black-box processor.

Without loss of generality, the kernel can always be re-expressed in terms of relative coordinates, $X \equiv (x_i - x')/2$, and average coordinates $(x_i + x')/2$: $K((x_i - x')/2, (x_i + x')/2; t, t')$. However, in many species, K has a weak dependence on the average coordinates. In other words, the kernel changes gradually with eccentricity or with angular distance from centre of gaze. Also, after adaptation it is known to be only a function of the temporal difference $T = t - t'$. Thus, to a first approximation one can assume translation invariance and retain only the dependence on the relative coordinates, $K(x_i, x'; t, t') = K(x_i - x'; t - t')$. This is convenient since it enables us to define the retinal filter, $K(f, w)$, simply by Fourier transforming K

$$K(\boldsymbol{f}, w) = \int \mathrm{d}\boldsymbol{X}\, \mathrm{d}T \exp(-\mathrm{i}\boldsymbol{f} \cdot \boldsymbol{X} - \mathrm{i}wT) K(\boldsymbol{X}, T). \tag{5.2}$$

This is the object that is actually measured in experiments. Furthermore, by rotational symmetry it is only a function of $(|\boldsymbol{f}|, w)$. In experiments, a luminosity grating, $L = I_0(1 + m\cos(fx)\cos(wt))$ is projected onto the RF of a cell and the minimum contrast m_{f,w,I_0} needed to elicit a certain level of response, r_0, at that spatio-temporal frequency of stimulation is recorded. The recording is repeated for different values of (f, w, I_0). By linearity of the output:

$$I_0 K_{I_0}(|\boldsymbol{f}|, w) = \frac{r_0}{m_{|\boldsymbol{f}|,w,I_0}}. \tag{5.3}$$

Thus there is a *family of retinal filters*, one for each adaptation or luminance level I_0. In figure 4, which is reproduced from the data of Enroth-Cugell and Robson (1966) and De Valois *et al* (1974), we show two typical families of filters, one for the cat and one for the monkey, as a function of f and at a given low temporal frequency w. More precisely, what is shown is $I_0 \times K$ which is called *contrast sensitivity*. A

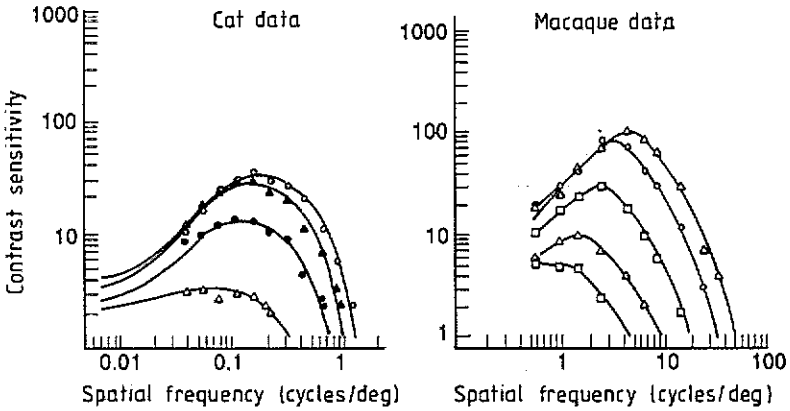

Figure 4. Measured contrast sensitivity. The data in the left figure are reproduced from Enroth-Cugell and Robson (1966), while that on the right are from De Valois *et al* (1974). In both cases, the luminance level I_0 decreases by one log unit each time we go to a lower curve.

prominent feature in that figure is the transition from band-pass to low-pass filtering as I_0 is lowered. A similar transition is also observed as the temporal frequency of stimulation is increased for a given spatial frequency.

If a retinal filter at high luminance is Fourier transformed back into space, it looks like the curve in figure 5. This is a one-dimensional slice in a two-dimensional rotationally invariant spatial profile, and it shows the familiar centre-surround organization of ganglion cell RF: The cell effectively receives excitatory input (+) from the photoreceptors in a small region around its RF centre and inhibitory input (−) from the surround region. These cells are known as *on*-centre cells. The other class of spatially opponent cells found in the retina have an inhibitory centre and an excitatory surround and are known as *off*-centre cells. A similar organization exists in the temporal domain.

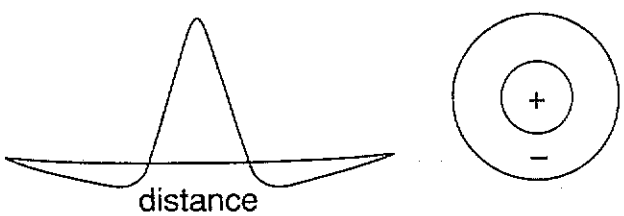

Figure 5. Retinal kernel at high adaptation level showing the opponent spatial organization of a ganglion cell's RF.

In retinas of species that possess colour vision, such as most primates and shallow-water fish, RFs of ganglion cells possess a more complicated centre-surround organization. In these retinas, there are several types of photoreceptors that possess different photosensitive pigments. Functionally, the various pigments are identical except they differ in the location of their peak spectral sensitivity. In humans for example, the three types of pigments referred to as B, G and R for blue, green, and red respectively (or alternatively known as S, M and L for short, medium and long spectral wavelength respectively) best absorb light of spectral wavelength around 419 nm, 530 nm and 558 nm respectively.

Corresponding to the diversity of photoreceptors there are several types of spatially opponent ganglion cells. These cells differ in the way the three photoreceptor types are used in the organization of their RF. In the primate retina, the most common on-centre ganglion cells receive excitatory input dominantly from one type of photoreceptors in the centre and inhibitory input from a different type in the surround: the two most common on-centre cell types are +R in centre and −G in the surround or +G in centre and −R in surround (Derrington *et al* 1984), see figure 6(*a*). The off-centre cells are similar. These colour coded cells are known as *single opponent* cells†.

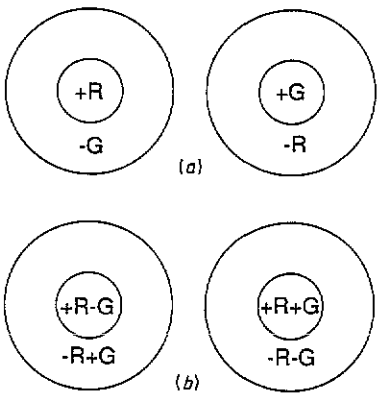

Figure 6. The two extremes of opponent colour coding. The cell types found in primates are shown in (*a*), while those found in shallow-water fish are shown in (*b*).

Single opponent cells are not found in retinas of all species that possess colour vision. In fact, they represent one extreme in colour coding. The other extreme is found in shallow-water fish which possess what are called *double opponent* cells (Daw 1968). As the name implies, these cells receive inputs of comparable strengths from two types of cones at every spatial location in their RF. For example, in one double opponent cell type found in goldfish retina, the RF has a centre that receives excitatory R and inhibitory G stimulation while its surround receives inhibitory R and excitatory G inputs, figure 6(*b*).

The fact that colour coding is qualitatively dependent on the environment of the animal makes it an interesting dimension for testing ecological theories. A successful theory of the retina should not only explain the shape of the retinal kernel and its dependence on background luminance, it should also account for differences seen among species. In the theory of retinal processing presented in subsections 5.2 and 5.3, differences in computation strategies among species are attributed to identifiable differences in the visual environment (information source differences). In subsection 5.2, we start by examining the problem in the purely spatial domain, and then show how to incorporate time. We also discuss something that we have ignored thus far, namely the role of noise. We introduce colour in subsection 5.3. The problem in pure space–time was first considered by Atick and Redlich (1990a,b), in the pure colour

† There are other opponent cell types that involve blue cones. However, since blue cones are rare in the retina (non-existent in fovea) these cells are also rare and hence will not be discussed here (De Monasterio *et al* 1985).

domain by Buchsbaum and Gottschalk (1983), and in the fully mixed dimension of space–time and colour by Atick *et al* (1990, 1991).

5.2. Theoretical approach to the retina: spatial processing

In section 3, we have given several reasons why a sensory pathway, such as the visual pathway, might recode incoming signals from the natural environment into a more efficient representation. In this section, we show how to use this idea to predict retinal processing in the spatial domain. We work with the hypothesis that the retina's main goal is to build a minimum entropy representation, i.e. a representation where the elements are statistically independent or decorrelated (the same procedure followed by the appropriate gain control yields a redundancy reduced code as we have seen in subsection 4.2.) However, we limit the class of recodings to linear transformations. Actually, with this restriction we shall see that the retina can only eliminate pairwise correlations†.

5.2.1. Decorrelation in the absence of noise. In subsection 3.2 the problem of finding minimum entropy codes was formulated as a variational problem of some well defined energy functional (5.3). The solution to the variational equations (3.3) then gives the optimal transformation that best minimizes (3.3). Here, we explicitly solve (3.3) for the class of linear mappings which, as discussed in the previous section, is the class to which the measured retinal transform belongs. We will also restrict this class further to that of one to one mappings purely for simplicity. The analysis can be repeated allowing the number of outputs to differ from the number of inputs—which is the biologicaly more realistic situation—however, our prediction for the organization of the receptive fields is insensitive to this assumption. With these simplifications (5.3) takes the following form

$$\begin{aligned} E\{\mathbf{K}\} &= \sum_{i=1}^{l} H(O_i) - 2\rho[H(\mathbf{O}) - H(\mathbf{L})] \\ &= \sum_{i=1}^{l} H(O_i) - \rho \log\det \mathbf{K}^{\mathrm{T}} \cdot \mathbf{K} \end{aligned} \tag{5.4}$$

where $O_i \equiv O(\boldsymbol{x}_i)$ is the response level of ganglion cell at location $\boldsymbol{x}_i$, and we have used the upright bold-face symbols to denote matrices and vectors; K denotes the matrix $K_{ij} \equiv K(\boldsymbol{x}_i - \boldsymbol{x}_j)$, $\mathbf{O} \equiv (O_1, \ldots, O_l)$, and similarly for L. We have also used the fact that $H(\mathbf{O}) - H(\mathbf{L}) = \log\det \mathbf{K} = \frac{1}{2}\log\det \mathbf{K}^{\mathrm{T}} \cdot \mathbf{K}$ which is valid when O is related to L through a linear transformation‡.

To exhibit $E\{\mathbf{K}\}$ more explicitly we need to compute the sum over pixel entropies $\sum_{i=1}^{l} H(O_i)$. Treating the discrete response levels O_i as a continuous variable, the ith pixel entropy can be approximated by a simple integral:

$$H(O_i) \equiv -\sum_{O_i} P(O_i)\log P(O_i) \to -\int \mathrm{d}O_i\, P(O_i)\log P(O_i) \tag{5.5}$$

† Since we will be assuming Gaussian signals, two-point decorrelation and statistical independence are equivalent.

‡ To see this, note that under a linear transformation $\mathbf{O} = \mathbf{K} \cdot \mathbf{L}$, the probabilities being densities—$\mathrm{d}\mathbf{O}P(\mathbf{O}) = \mathrm{d}\mathbf{L}P(\mathbf{L})$ transform as $P(\mathbf{O}) = P(\mathbf{L})/\det \mathbf{K}$. Substituting this expression into the definition of $H(\mathbf{O})$ and changing variables it is straightforward to get $H(\mathbf{O}) = H(\mathbf{L}) + \log\det \mathbf{K}$.

which depends on the ith pixel probability $P(O_i)$. The latter is computable from the input probability $P(\mathbf{L})$ since $O_i = \sum_{j=1}^{l} K_{ij} L_j$. $P(\mathbf{L})$ in (3.7) is a Gaussian with a covariance matrix $\mathbf{R}$, therefore $P(\mathbf{O})$ is also a Gaussian but with covariance matrix $\tilde{\mathbf{R}} \equiv \mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^{\mathrm{T}}$. In (5.5) we only need the individual pixel probability $P(O_i)$ for every i, which is given by

$$P(O_i) = \int \prod_{j \neq i} \mathrm{d}O_j \, P(\mathbf{O}).$$

It is easy to do the integrals and show that

$$P(O_i) = \frac{1}{2\pi \tilde{R}_{ii}} \exp\left(-\frac{1}{2\tilde{R}_{ii}} O_i^2\right). \tag{5.6}$$

(Again $\tilde{R}_{ii} = \langle O_i^2 \rangle$, the diagonal part of $\tilde{R}_{ij} = \langle O_i O_j \rangle$.)

Substituting the expression for $P(O_i)$ from (5.6) in (5.5), we find $H(O_i) = \log \tilde{R}_{ii}$, which when summed over all pixels yields

$$\sum_{i=1}^{l} H(O_i) = \log \prod_{i=1}^{l} \tilde{R}_{ii}. \tag{5.7}$$

By translation invariance, all the $\tilde{R}_{ii}$ are equal, $\tilde{R}_{ii} = \langle O_0^2 \rangle$ for a pixel at some arbitrary location 0, thus $\sum_{i=1}^{l} H(O_i) = l \log(\langle O_0^2 \rangle)$. This can be substituted for the first term in (5.4); however, there are a couple of mathematical steps that lead to an even simpler form of the energy functional.

First since $\langle O_0^2 \rangle \geqslant 0$, we can drop the logarithm from $\log(\langle O_0^2 \rangle)$ and minimize instead the simpler quantity $\langle O_0^2 \rangle$. However, by translation invariance, minimizing $\langle O_0^2 \rangle$ is equivalent to minimizing the explicitly invariant expression $\sum_i \langle O^2(\boldsymbol{x}_i) \rangle = \sum_i (\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^{\mathrm{T}})_{ii} = \mathrm{Tr}(\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^{\mathrm{T}})$. The final energy function is then

$$E\{\mathbf{K}\} = \mathrm{Tr}(\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^{\mathrm{T}}) - \rho \log \det(\mathbf{K}^{\mathrm{T}} \cdot \mathbf{K}). \tag{5.8}$$

The advantage of this invariant form of E is that we can now go to Fourier space very easily:

$$E\{K\} = \int \mathrm{d}f |K(f)|^2 R(f) - \rho \int \mathrm{d}f \log |K(f)|^2 \tag{5.9}$$

where we have used the identity $\log \det \mathbf{Q} = \mathrm{Tr} \log \mathbf{Q}$ valid for any positive matrix $\mathbf{Q}$.

The variational equations in frequency space, $\delta E\{K\}/\delta K(f) = 0$, are trivial in this case: the optimal solution is just

$$|K(f)|^2 = \frac{\rho}{R(f)}. \tag{5.10}$$

This could have been guessed more easily by diagonalizing the autocorrelator matrix of the output $\tilde{R}(\boldsymbol{x}_i - \boldsymbol{x}_j) \equiv \langle O(\boldsymbol{x}_i) O(\boldsymbol{x}_j) \rangle$. However, we have gone through the

analysis systematically to illustrate the general procedure which will be useful for more complex codings. Since $R(f) = 1/|f|^2$ for natural scenes, the predicted kernel is simply $K(f) = \rho|f|$. On a log–log plot this gives a curve of slope one.

We can compare this simple prediction with retinal filters in the regime where the noise is not significant, namely in the regime of high luminance I_0 and at low frequencies. In figures 7(*a*) and 7(*c*) we have plotted some typical experimentally measured retinal filters at high luminance I_0. The data are taken from De Valois *et al* (1974) and from Kelly (1972), respectively. In figures 7(*b*) and 7(*d*), we show the ratio $\chi(f) = K_m(f)/K_p(f)$ where K_m and $K_p \sim |f|$ are the measured and predicted filters respectively. At low frequency, we can see that $\chi(f)$ is flat or that both filters have the same slope.

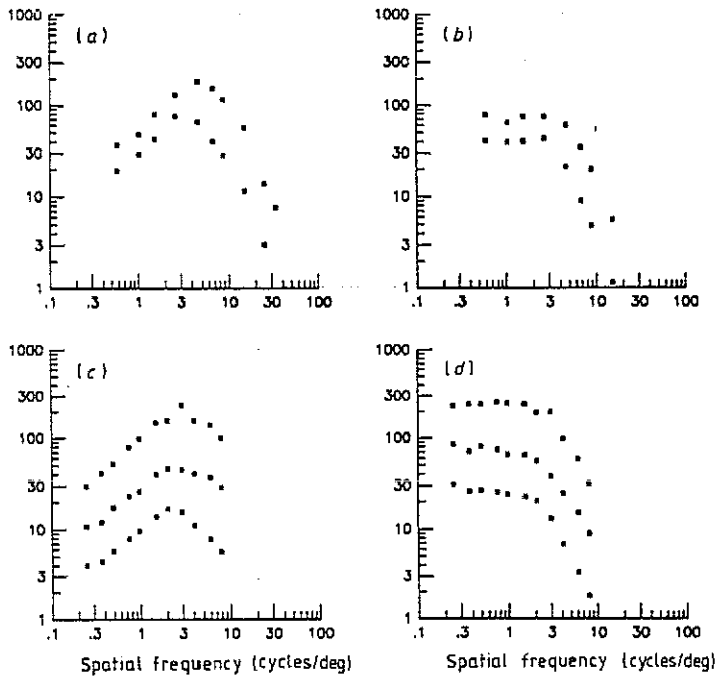

Figure 7. Retinal filters (*a*), and (*c*) at high mean luminosities, taken from the data of De Valois *et al* (1974) and Kelly (1972), respectively. (*b*) and (*d*) are the data in (*a*) and (*c*), respectively, multiplied by $1/|f|$, which is the amplitude spectrum of natural scenes. This gives the retinal ganglion cell's output amplitude spectrum. Notice the whitening of the output at low frequency. The ordinate units are arbitrary.

Another way to interpret the results in figures 7(*b*) and 7(*d*) is as follows. The power spectrum of the output is given by the square of the retinal filter times the input power spectrum:

$$\langle O(f)O^*(f)\rangle = \langle (K(f)L(f))(K(f)L(f))^*\rangle = |K(f)|^2 R(f). \tag{5.11}$$

However, $R(f)$ from (3.6) is $1/|f|^2$. The output amplitude spectrum, which is the square root of the power spectrum, is then proportional to $\chi(f)$ which is what is plotted in figures 7(*b*) and 7(*d*). Thus at low frequencies, the input spectrum $|f|^{-2}$ is converted by the retinal kernel $K(f)$ into a flat spectrum at the retinal output:

$\langle O(f)O^*(f)\rangle$ = constant. This *whitening* of the input by the retina continues up to the frequency where the kernel peaks in figures 7(*a*) and 7(*c*). Beyond that the noise is no longer ignorable and the actual kernel deviates from the pure whitening kernel. This whitening is the statement in frequency space of decorrelation in regular space. Of course, since the whitening does not continue all the way to the system's cutoff the decorrelation in space is not perfect. In the next section we shall see that by incorporating a strategy for noise suppression in addition to decorrelation we arrive at filters that agree with what is measured not only over the entire range of visible spatial frequencies but also at all luminance levels.

5.2.2. Decorrelation in the presence of noise. The above agreement does support a strategy of decorrelation in the absence of noise. However, decorrelation cannot be the only goal in the presence of input noise such as photon (or quantum) noise which always exists. In that case, decorrelation alone would be a dangerous computational strategy as we now argue: If the retina were to whiten all the way up to the cutoff frequency or resolution limit, the kernel $K(f)$ would be proportional to $|f|$ up to that limit. This would imply a constant average squared response KRK^* to natural signals $L(x)$, which for $R \sim |f|^{-2}$ have large spatial power at low frequencies and low power at high frequencies. But this same $K(f) \sim |f|$ acting on input noise whose spatial power spectrum is approximately flat has a very undesirable effect, since it amplifies the noise at high frequencies where noise power, unlike signal power, is not becoming small. Therefore, even if input noise is not a major problem without decorrelation, after complete decorrelation (or whitening up to cutoff) it would become a problem. Also, if both noise and signal are decorrelated at the output, it is no longer possible to distinguish them. Thus, if decorrelation is a strategy, there must be some guarantee that no significant input noise is passed through the retina to the next stage. We believe this is why the retina stops whitening its input at a frequency far lower that the cutoff frequency.

Further evidence that the retina is concerned about not passing significant amounts of input noise is found in the fact that the ganglion cell kernel, as we have seen in subsection 5.1, makes a transition from band-pass to complete low-pass as the retina adapts to very low I_0. Since as I_0 decreases the signal to noise ratio of the input signals decreases, one expects low-pass filtering as a way of suppressing the noise, which is what the retina does.

Since here we are primarily interested in testing the predictions of minimum entropy coding (equivalently redundancy reduction), we take a somewhat simplified approach to the problem with noise. Instead of doing a full-fledged information theoretic analysis that unifies minimum entropy with noise suppression (as in Atick and Redlich 1990a,b), we work in a formalism where the signal is first low-pass filtered to eliminate noise and the resulting signal is then decorrelated as before. The advantage of this modular approach is that it leads to a more intuitive picture of the various processing stages in the retina and it also gives parameters that have physical significance. Furthermore, the analysis is not as complicated as that in the unified formalism.

We start by going over the stages of signal processing that we assume precede the decorrelation stage. Figure 8 shows a schematic of those stages. First, images from natural scenes pass through the optical medium of the eye and in doing so their image quality is lowered. It is well known that this effect can be taken into account by multiplying the images by the optical *modulation transfer function* or MTF of the eye,

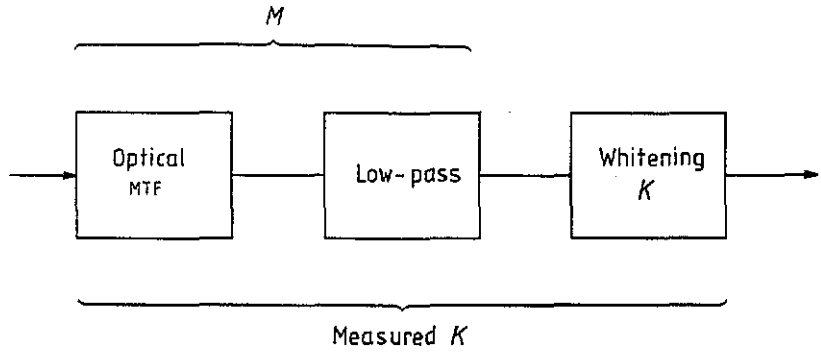

Figure 8. Schematic of the signal processing stages assumed to take place in the retina.

a function of spatial frequency that is measurable in purely non-neural experiments (Campbell and Gubisch 1966). In fact, an exponential of the form $\exp(-(|f|/f_c)^\alpha)$, for some scale f_c characteristic of the animal (in primates $f_c \sim 22$ cycles/deg and $\alpha \sim 1.4$) is a good approximation to the optical MTF. The resulting image is then transduced by the photoreceptors and is low-pass filtered to eliminate input noise. Finally, we assume that it is decorrelated. In this model, the output–input relation schematically takes the form

$$O = K \cdot (M \cdot (L + n) + n_0) \tag{5.12}$$

where the dot denotes a convolution as defined in (5.1), $n(x)$ is the input noise (such as quantum noise) while $n_0(x_i)$ is some intrinsic noise level which models post-receptor synaptic noise. Finally, M is the filter that takes into account both the optical MTF as well as the low-pass filtering needed to eliminate noise. An explicit expression for M will be derived below.

With this model, the energy functional determining the decorrelation filter $K(f)$ is

$$E\{K\} = \int \mathrm{d}f\ |K(f)|^2\,[M^2(f)(R(f) + N^2) + N_0^2] - \rho \int \mathrm{d}f \log |K(f)|^2 \tag{5.13}$$

where $N^2(f) \equiv \langle |n(f)|^2 \rangle$ and $N_0^2(f) \equiv \langle |n_0(f)|^2 \rangle$ are the input and synaptic noise powers respectively. This energy functional is the same as that in (5.9) but with the variance $\tilde{R}(f)$ replaced by the variance of O in (5.12).

As before, the variational equations $\delta E/\delta K(f) = 0$, are easy to solve for $K(f)$. The predicted filter that should be compared with experimental measurements is this variational solution, K, times the filter M. We denote this by K_{expt}:

$$|K_{\text{expt}}(f)| = |K(f)|\ M(f) = \frac{\sqrt{\rho} M(f)}{[M^2(f)\,(R(f) + N^2) + N_0^2]^{1/2}}. \tag{5.14}$$

An identical result can be obtained in space–time trivially by replacing the autocorrelator $R(f)$ and the filter $M(f)$ by their space–time analogues $R(f, w)$ and $M(f, w)$, respectively, with w the temporal frequency. However, we focus here on the purely spatial problem where we have Field's (1987) measurement of the spatial autocorrelator $R(f)$ of natural scenes: $R(f) = I_0^2/|f|^2$.

5.2.3. Deriving the low-pass filter. In our explicit expression for K_{expt}, below, we shall use the following low-pass filter

$$M(f) = \frac{1}{N}\left(\frac{1}{I_0}\frac{R(f)}{R(f)+N^2}\right)^{1/2}\exp\left[-\left(\frac{|f|}{f_c}\right)^{\alpha}\right]. \tag{5.15}$$

The exponential term is the optical MTF while the first term is a low-pass filter that we derive next using information theory. The reader who is not interested in the details of the derivation can skip this rather technical section without loss of continuity.

It is not clear in the retina what principle dictates the choice of the low-pass filter or how much of the details of the low-pass filter influence the final result. In the absence of any strong experimental hints, of the type which imply redundancy reduction, we shall try a simple information theoretic principle to derive an M. We insist that the filter M should be chosen such that the filtered signal $O' = M\cdot(L+n)$ carries as much information as possible about the *ideal* signal L subject to some constraint. To be more explicit, the amount of information carried by O', about L, is the mutual information $I(O', L)$ (see the appendix). However, as we discuss in the appendix $I(O', L) = H(O')$ – noise entropy (for L and n statistically independent Gaussian variables), and thus if we maximize $I(O', L)$ keeping fixed the entropy $H(O')$ we achieve a form of noise suppression.

We can now formulate this as a variational principle. To simplify the calculation we assume Gaussian statistics for all the stochastic variables involved. The output–input relation takes the form: $O' = M\cdot(L+n)+n_0$. A standard calculation leads to

$$I(O', L) = \int \mathrm{d}f \log\left(\frac{M^2(R+N^2)+1}{M^2N^2+1}\right) \tag{5.16}$$

where we have chosen units where the quantization noise have unit variance $\langle n_0^2\rangle = 1$. Similarly, one finds for the entropy $H(O') = \int \mathrm{d}f \log(M^2(R+N^2)+1)$ in the same units. The variational functional or energy for smoothing can then be written as $E\{M\} = -I(O', L) + \eta H(O')$. It is not difficult to show that the optimal noise suppressing solution $\delta E/\delta M = 0$ takes the form:

$$M = \left(\frac{1}{N}\right)\left(\frac{1}{\eta}\frac{R}{R+N^2}-1\right)^{1/2}. \tag{5.17}$$

If the parameter $\eta \geqslant 1$ then clearly there is no non-vanishing solution M to this smoothing problem. We will assume that $\eta \ll 1$ so in fact we are in the regime where the first term inside the square root dominates and hence we can drop the -1 term. Actually η has a dependence on I_0 since to hold $H(O')$ fixed at all values of I_0 implies that η be a function of I_0. It is not hard to see that $\eta \sim I_0$ will ensure that $H(O')$ fixed with mean luminance (we assume that noise N^2 is quantum noise, and hence $N^2 \sim I_0$). Ignoring all overall factors in M that are independent of f and I_0 we arrive at the expression that we exhibit in the first term in (5.15).

5.2.4. Analysing the solution. Let us now analyse the form of the complete solution (5.14), with M given in (5.15). In figure 9 we have plotted $K_{\text{expt}}(f)$ (curve A) for a typical set of parameters. We have also plotted the filter without noise $R(f)^{-1/2}$,

(5.10), (curve B) and $M(f)$, (5.15), (curve C). There are two points to note: at low frequency the kernel $K_{\text{expt}}(f)$ (curve A) is identically performing decorrelation, and thus its shape in that regime is completely determined by the statistics of natural scenes: the physiological functions M and N drop out. At high frequencies, on the other hand, the kernel coincides with the function M, and the power spectrum of natural scenes R drops out.

We can also study the behaviour of the kernel in (5.14) as a function of mean luminosity I_0. If one assumes that the dominant source of noise is quantum noise, then the dependence of the noise parameter on I_0 is simply $N^2 = I_0 N'^2$ where N' is a constant independent of I_0 and independent of frequency (flat spectrum). This gives an interesting result. At low frequency where K_{expt} goes like $1/\sqrt{R}$ and its I_0 dependence will be $K_{\text{expt}} \sim 1/I_0$ (recall $R \sim I_0^2$), the system exhibits a Weber law behaviour, i.e. its contrast sensitivity $I_0 K_{\text{expt}}$ is independent of I_0. In the other regime—at high frequency—where the kernel asymptotes to M with $N^2 \gg R$, then $K_{\text{expt}} \sim 1/I_0^{1/2}$ which is a De Vries–Rose behaviour $I_0 K_{\text{expt}} \sim I_0^{1/2}$. This predicted transition from Weber to De Vries–Rose with increasing frequency is in agreement with what is generally found (see Kelly 1972, figure 3).

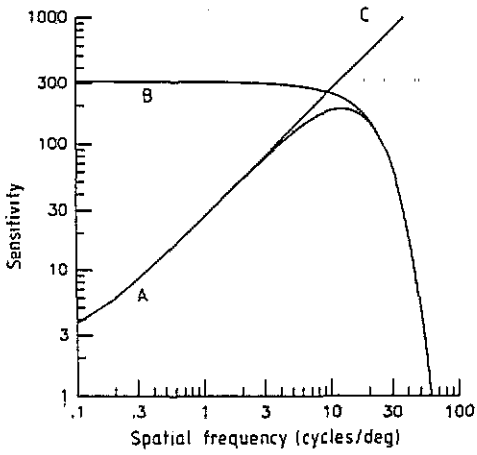

Figure 9. *A typical predicted retinal filter (curve* A) from (5.14), while curve C is $R(f)^{-1/2}$ which is the pure whitening filter (5.10). Finally curve B is the low-pass filter M.

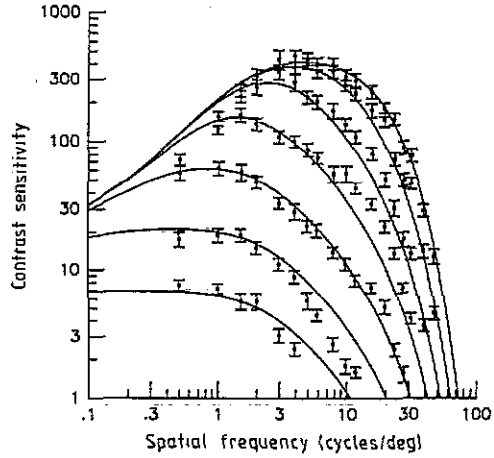

Figure 10. *Predicted retinal filters, (5.14), at different* I_0 separated by one log units, assuming that the dominant source of input noise is quantum noise ($N^2 \sim I_0$). No other parameters depend on I_0. The fixed parameters are $f_c = 22$ cycles/deg, $\alpha = 1.4$, $\rho = 2.7 \times 10^5$, $N' = .075$. The data are psychophysical contrast sensitivity measurements of Van Nes and Bouman (1967).

Given the explicit expression in (5.14) and the choice of quantum noise for N we can generate a set of kernels as a function of I_0. The resulting family is shown for primates in figure 10. We need to emphasize that there are no free parameters here which depend on I_0. The variables that needed to be fixed were the numbers f_c, α, ρ, N' and N_0 and they are independent of I_0. Also we work in units of synaptic noise n_0, so the synaptic noise power N_0^2 is set to one. We have superimposed on this family the data from the experiments of Van Nes and Bouman (1967) on human

psychophysical contrast sensitivity. It does not take much imagination to see that the agreement is very reasonable especially keeping in mind that this is not a fit but a *parameter-free* prediction.

We can also compare the predicted kernels in (5.14) to ganglion cell kernels measured in single-cell experiments. However *a priori* we do not expect close quantitative agreement for the following reason. The theory predicts the transfer function for the aggregate system, i.e. for the collection of all the output cells. The only reason why it might appear that we predicted a single-cell kernel is because of the assumption of translation invariance which forced all cells to have the same kernel and forced single-cell properties to be identical to aggregate properties. In the real retina, where translation symmetry is broken, different cells have different kernels and hence the aggregate transfer function (obtained by combining the different single-cell kernels) is not equal to any one single-cell kernel. In experiments, it turns out that the difference between single-cell and psychophysics data is quantitative not qualitative. For instance, in the psychophysical data one finds that the frequency of optimal contrast sensitivity decreases with I_0 more than what is found in single-cell experiments. The shifts that the theory predicts are more consistent with the psychophysical shifts. However, we expect that repeating the calculations of the last section without translation invariance would produce a family of single-cell kernels with shifts that are in closer quantitative agreement with experiment.

5.3. Introducing colour

Images in nature carry information through their spectral compositions in addition to their spatio-temporal modulations. So an image is generally a function of the form $L(\boldsymbol{x}, t, \lambda)$, where λ is the spectral wavelength. Many animals have evolved visual pathways capable of extracting this colour information. In the retina of these species, images are first sampled in the spectral domain through the three cone types to give the output activities

$$P^a(\boldsymbol{x}, t) = \int \mathrm{d}\lambda C^a(\lambda) L(\boldsymbol{x}, t, \lambda) + n(\boldsymbol{x}, t) \tag{5.18}$$

where the functions $C^a(\lambda)$ are the spectral sensitivity functions for the three photoreceptor types, $a = 1, 2, 3$ for R, G and B respectively. In figures 11(a) and 11(b) we show the spectral sensitivity curves for the cones in the retinas of primates and shallow-water fish, respectively (the two systems that form the two extremes in retinal colour coding).

One important feature to notice about the two sets of curves in figure 11 is the fact that the R and G spectral sensitivity curves overlap. The degree of overlap is more significant for primates' retina than for shallow-water fish. To be quantitative, in the monkey *Macaca fascicularis* the separation of the spectral peak sensitivities between R and G is about 30 nm while for goldfish the corresponding separation is about 90 nm. This difference between the two species is due to adaptation of cone pigments to different visual environments† and will play an essential role in explaining

† In the case of primates, which are believed to have evolved in a forest like environment, one finds that the proximity of R and G cones can be explained by the fact that most of the information in a forest is squeezed in a narrow spectral band centred about 550 nm. Thus one needs to sample that region more densely if one is to resolve different objects found in that spectral band. On the other hand,

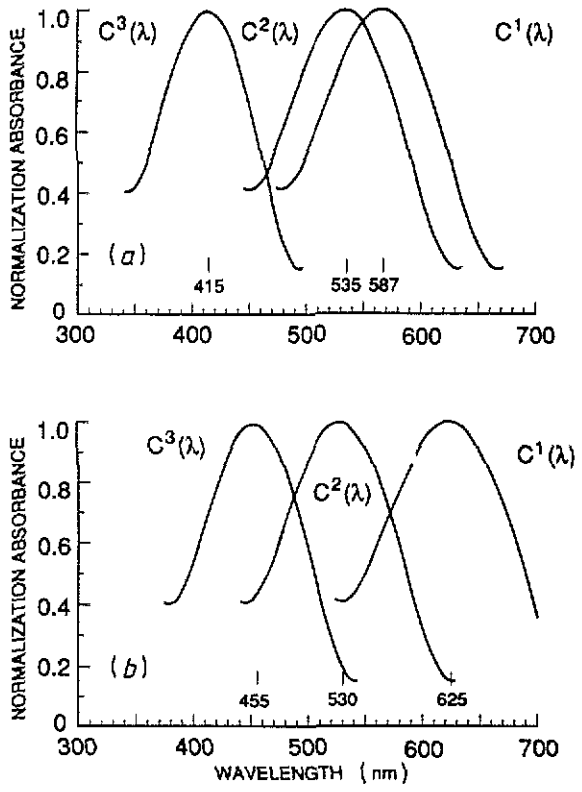

Figure 11. Spectral sensitivity curves for primates (*a*), and shallow-water fish, (*b*). Notice that the overlap of C^1 and C^2 expecially in the primates.

the subsequent differences in the neural coding strategies for colour mentioned in subsection 5.1.

The fact that the cones sample in such an overlapping fashion introduces an additional source of correlations in the photoreceptor signals and thus an additional source of inefficiency that has to be eliminated (Barlow 1961, Buchsbaum and Gottschalk 1983, Atick *et al* 1990). We will limit our analysis to the two-cone (R and G) system, since in primate retina these photoreceptors occur with equal density and are more abundant than the blue cones. In fact, the blue cones constitute only 15% of the total cone population in the entire retina while in the fovea they are virtually non-existent. For a discussion of the role of blue cones see Atick *et al* (1990).

We now generalize the analysis of the previous subsection 5.2 to include colour. The chromatic-spatio-temporal correlator is a matrix of the form $R^{ab}(\boldsymbol{x},t)$ or in Fourier space $R^{ab}(\boldsymbol{f},w)$. Here $R^{11}(\boldsymbol{x},t)$ is the red–red correlator $\langle P^1(0,0)P^1(\boldsymbol{x},t)\rangle$ and $R^{12}(\boldsymbol{x},t)$ is the red–green correlator defined similarly and so on. Unfortunately, not much is known experimentally about the entries of this

under water light in the spectral band between 550 nm and 610 nm is heavily absorbed by water with the amount of absorption increasing dramatically with distance travelled. Thus if shallow-water fish had adopted pigments around 568 nm just like primates, they would not have been able to see far under water. Shallow-water fish instead evolved cones that sampled near the infrared, an area where the signal under water travels much farther before complete absorption. Additional discussion regarding the adaptation of the cone system of various species to the environment can be found in the excellent book of Lythgoe (1979).

matrix. Thus, we are forced to make some assumptions. Although, it is possible to do the analysis entirely for the most general form of $R^{ab}(f,w)$ (see Atick *et al* 1990), it is just as informative and much simpler to analyse the case where $R^{ab}(f,w)$ can be factorized into a pure spatio-temporal correlator times a 2×2 matrix describing the degree of overlap between the R and G systems. We will also only examine colour coding under conditions of slow temporal stimulation or zero temporal frequency. In that case, we can replace the spatio-temporal correlator by by $I_0^2/|f|^2$ (3.6). Thus we take

$$R^{ab}(f,0) = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \frac{I_0^2}{|f|^2} \tag{5.19}$$

where $r < 1$ is a parameter describing the degree of overlap of R and G. We should emphasize, that we do not advocate that this is the form of R^{ab} necessarily found in nature. We have reduced R^{ab} to one degree of freedom in order to illustrate very simply the possibilities. More complex R^{ab}, in particular those where space and colour are not decoupled, lead to quantitatively different but qualitatively similar solutions.

As before, the output **O** is related to the input **P** through

$$\mathbf{O} = \mathbf{K} \cdot (\mathbf{M} \cdot (\mathbf{P} + \mathbf{n}) + \mathbf{n}_0) \tag{5.20}$$

where $n^a(x,t)$ is input noise including transduction and quantum noise, while $n_0(x,t)$ is noise (e.g. synaptic) added following the low-pass filter **M**. We have introduced upright bold face to denote in this section matrices in the 2×2 colour space; also in (5.20) each $\cdot$ denotes a convolution in space. To see how the presence of two channels affect the spatial low-pass filtering, it is helpful to rotate in colour space to the basis where the colour matrix is diagonal. For the simple colour matrix in (5.19), this is a 45 degree rotation from the red **R** and green **G** basis to the luminance, **G+R**, and chromatic, **G − R**, channels (in vector notation, the red and green channels are denoted by $\mathbf{R} = (\mathbf{1},\mathbf{0})$ and $\mathbf{G} = (\mathbf{0},\mathbf{1})$). This 45 degree rotation matrix is

$$\mathbf{U}_{45} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{5.21}$$

In the $\mathbf{G} \pm \mathbf{R}$ basis, the total correlation matrix plus the contribution due to noise is

$$\mathbf{U}_{45}\,(\mathbf{R}(f) + \mathbf{N}^2)\,\mathbf{U}_{45}^{\mathrm{T}} = \frac{I_0^2}{|f|^2} \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} + N^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5.22}$$

where the noise, $\langle n^a n^b \rangle = \delta^{ab} N^2$, is assumed equal in both the **R** and **G** channels, for simplicity. Since in the $\mathbf{G} \pm \mathbf{R}$ basis the two channels are decoupled, the spatial filters $M_\pm(f)$ are found by applying our single-channel result in (5.15). More specifically they are found by replacing $R(f)$ in (5.15) by

$$R_\pm(f) = (1 \pm r) I_0^2/|f|^2. \tag{5.23}$$

Notice that the two channels differ only in their effective signal-to-noise ratios: $(S/N)_\pm = \sqrt{(1 \pm r)}(I_0/N)$ which depend multiplicatively on the colour eigenvalues

$1 \pm r$. In the luminance channel, $\mathbf{G} + \mathbf{R}$, the signal-to-noise is increased above that in either the **R** or **G** channel alone, due to the summation over the **R** and **G** signals. The filter $M_+(f)$, therefore, passes relatively higher spatial frequencies, increasing spatial resolution, than without the **R** plus **G** summation. On the other hand, the chromatic channel, $\mathbf{G} - \mathbf{R}$, has lower S/N, proportional to $1 - r$, so its spatial filter $M_-(f)$ cuts out higher spatial frequencies, thus sacrificing resolution in favour of colour discriminability. The complete filter is finally obtained by rotating from the $\mathbf{G} \pm \mathbf{R}$ basis by 45 degrees back to the **R** and **G** basis

$$M^{ab}(f) = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} M_+(f) & 0 \\ 0 & M_-(f) \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{5.24}$$

(Again $M_\pm(f)$ is given by (5.15) with $R(f) \to R_\pm(f)$ in (5.23).)

After filtering noise, the next step is to decorrelate the signal as if no noise existed as we did in the purely spatial problem (subsection 5.2). In this case this means that we have to find the kernel **K** that achieves diagonalization of the low-pass filtered spatio-chromatic autocorrelator

$$\hat{\mathbf{R}} = \langle(\mathbf{M}\cdot(\mathbf{P}+\mathbf{n}) + \mathbf{n}_0)\cdot(\mathbf{M}\cdot(\mathbf{P}+\mathbf{n}) + \mathbf{n}_0)^{\mathrm{T}}\rangle. \tag{5.25}$$

In other words we need to find **K** that satisfies $\mathbf{K}\cdot\hat{\mathbf{R}}\cdot\mathbf{K}^{\mathrm{T}} = \mathbf{D}$ with **D** a diagonal matrix in space and colour. In the purely spatial problem, we have insisted on a translationally invariant, local set of retinal filters: the approximation where all retinal ganglion cells (in some local neighbourhood, at least) have the same receptive fields, except translated on the retina, and these fields sum from only a nearby set of photoreceptor inputs. These assumptions force **D** to be proportional to the unit matrix. In generalizing this to include colour, we note that when **D** is proportional to the unit matrix, the mean squared outputs ($(K\hat{R}K^{\mathrm{T}})^{aa}_{xx}$ for output O^a_x) of all ganglion cells are equal. This equalization provides efficient use of optic nerve cables (ganglion cell axons) if the set of cables for the cells in a local neighbourhood all have similar information carrying capacity. We therefore continue to take **D** proportional to the identity matrix in the combined space–colour system. Taking **D** proportional to the identity, however, still leaves a freedom to arbitrarily mix the proportion of the two decorrelated colour signals since one can still rotate by a 2×2 orthogonal matrix U^{ab}_θ, i.e. $\mathbf{K}(f) \to \mathbf{U}_\theta\mathbf{K}(f)$, that leaves **D** proportional to the identity†. This freedom to rotate by $\mathbf{U}_\theta$ will be eliminated later by looking at how much information (basically S/N) is carried by each channel. We shall insist that no optic nerves are wasted carrying signals with very low S/N.

We are now ready to write down the prediction for $K^{ab}(f)$. To do that we go to the $\mathbf{G} \pm \mathbf{R}$ basis where $M^{ab}(f)$ is diagonal in colour space. $K^{ab}(f)$ can then be taken to be diagonal since there are no correlations in colour in that basis: it consists of two functions $K_\pm(f)$ which are chosen to separately whiten the $\mathbf{G} \pm \mathbf{R}$ channels. Since the complete frequency space correlators in the two channels after filtering by $M_\pm(f)$ are $M^2_\pm(f)(R_\pm(f) + N^2) + N_0^2$, the $K_\pm(f)$ are therefore

$$K_\pm(f) = \frac{\sqrt{\rho}}{\left[M^2_\pm(f)(R_\pm(f) + N^2) + N_0^2\right]^{1/2}} \tag{5.26}$$

† U^{ab}_θ is a constant matrix depending only on one number, the rotation angle; it satisfies $\mathbf{U}_\theta\mathbf{U}^{\mathrm{T}}_\theta = 1$.

where N_0^2 is the power of the noise which is added following the filter $M^{ab}(f)$, see (5.20).

Now putting (5.26) together with (5.23) and (5.15), we obtain the complete retinal transfer function, i.e. the one to be compared with experiment,

$$\mathbf{K}_{\text{expt}} = \mathbf{U}_\theta \begin{pmatrix} K_+(f) & 0 \\ 0 & K_-(f) \end{pmatrix} \begin{pmatrix} M_+(f) & 0 \\ 0 & M_-(f) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{5.27}$$

As a reminder, the rightmost matrix transforms the **G**, **R** inputs into the $\mathbf{G} \pm \mathbf{R}$ basis. These signals are then separately filtered by $K_\pm M_\pm$. Finally, the rotation $\mathbf{U}_\theta$ to be specified shortly, determines the mix of these two channels carried by individual retinal ganglion cells. We should emphasize that the outputs of the two colour channels defined by (5.27) continue to be decorrelated for any choice of rotation angle θ.

5.3.1. Analysing the colour solutions. In this section we show how the diverse processing types such as those found in goldfish and primates are both given by (5.27) but for different values of the parameter r in the colour correlation matrix.

For the case of goldfish, as mentioned earlier, one expects only small overlap between R and G responses and thus r is small. The diagonal channels $\mathbf{G} \pm \mathbf{R}$ then have eigenvalues $1 \pm r$ of the same order: $(1 - r)/(1 + r) \sim 1$. This means both channels on average carry roughly the same amount of information and transmit signals of comparable S/N. Thus the filters $K_+(f)M_+(f)$ and $K_-(f)M_-(f)$ are very similar. In fact, they are both band-pass filters as shown in figure 12(*a*) for some typical set of parameters. Since these channels are already nearly equalized in S/N, there is no need to mix them by rotating with $\mathbf{U}_\theta$, so that matrix can be set to unity. Therefore, the complete solution (5.27) when acting on the input vectors $\mathbf{R}, \mathbf{G}$, gives two output channels corresponding to two ganglion cell types:

$$\begin{aligned} \mathbf{Z}_1 &= (\mathbf{G} + \mathbf{R})\ K_+ M_+ \\ \mathbf{Z}_2 &= (\mathbf{G} - \mathbf{R})\ K_- M_-. \end{aligned} \tag{5.28}$$

If we Fourier transform these solutions to get their profiles in space, we arrive at the kernels $K^{ab}(x - x')$ shown in figure 13 for some typical set of parameters. The top row is one cell type acting on the **R** and **G** signals, and the bottom row is another cell type. These have features of double opponency cells.

Moving to primates, there is one crucial difference which is the expectation that r is closer to 1 since the overlap of the spectral sensitivity curves of the red and green is much greater: the ratio of eigenvalues $(1 - r)/(1 + r) \ll 1$. Since the colour eigenvalues modify the S/N, this implies that the $\mathbf{G} - \mathbf{R}$ channel has a low S/N while the $\mathbf{G} + \mathbf{R}$ has much higher S/N. Therefore, $K_-(f)M_-(f)$ is a low-pass filter while $K_+(f)M_+(f)$ is band-pass as shown in figure 12(*b*). These two channels can be identified with the chromatic and luminance channels measured in psychophysical experiments, respectively. The curves shown in figure 12(*b*) do qualitatively match the results of psychophysical contrast sensitivity experiments (Mullen 1985): namely the low-pass and band-pass properties of the chromatic and luminance curves, respectively.

Although there is psychophysical evidence that indicates that colour information in primate cortex is organized into luminance and chromatic channels under normal

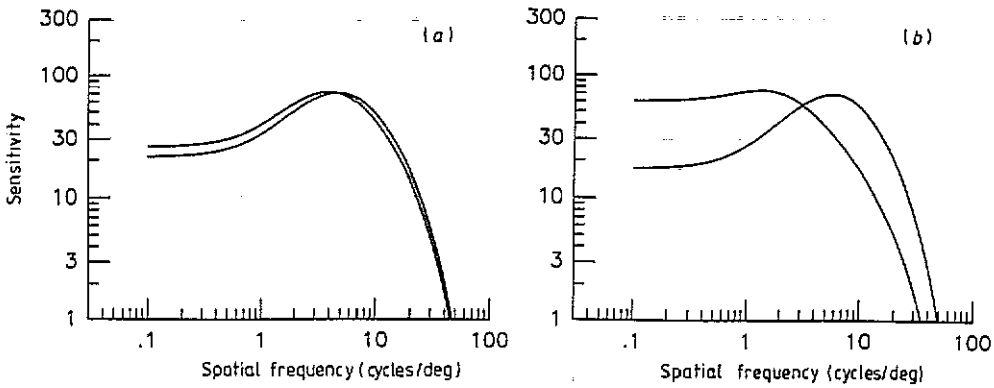

Figure 12. The luminance and chromatic channels for goldfish, (*a*), and for primates, (*b*). In both figures the curve that is more like band-pass is for the luminance $\mathbf{G}+\mathbf{R}$ channel, while the other is for the $\mathbf{G}-\mathbf{R}$ channel. Parameters used are $\bar{I}_0/\bar{N}=5.0$, $\alpha=1.4$, $f_c=22.0$ cycles/deg, $N_0=1.0$ for both (*a*) and (*b*), the differences being that $r=0.2$ for (*c*) and 0.85 for (*b*).

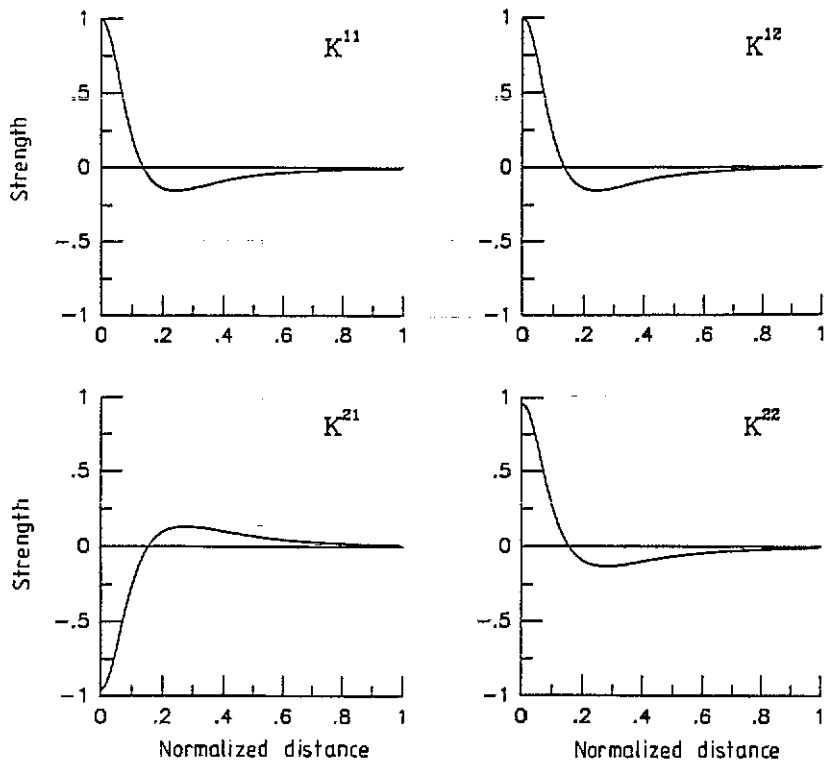

Figure 13. Predicted retinal kernel K^{ab} in the **R** and **G** basis in the goldfish regime $r=0.2$ and for the same parameters as those in figure 12. These cells can be termed double opponency cells.

adaptation conditions (Mullen 1985), this is not how the primate retina transmits information down the optic nerve (Derrington *et al* 1984). One reason why the primate retina may choose not to use the $\mathbf{G}\pm\mathbf{R}$ basis is that the representation of informa-

tion in chromatic and luminance channels has one undesirable consequence. If we compute the signal-to-noise ratio as a function of frequency in the chromatic channel, given by $(S/N)_-^2 = K_-^2 M_-^2 R_-/[K_-^2(M_-^2 N^2 + 1)]$, and compare it with the corresponding ratio in the luminance channel we find that the ratio $(S/N)_-/(S/N)_+ \ll 1$ because $(1-r)/(1+r) \ll 1$. So for primates, transmitting information in the luminance and chromatic basis would result in one channel with very low S/N, or equivalently one channel that does not carry much information. Transmitting information at low S/N down the optic nerve could be dangerous, especially since the optic nerve introduces intrinsic noise of its own; it also may be wasteful of optic nerve hardware. What we propose here is to use the remaining symmetry of multiplication by the rotation matrix $\mathbf{U}_\theta$ to mix the two channels so they carry the same amount of information, i.e. such that they have the same S/N at each frequency. Keep in mind that this does not affect the decorrelated nature of the two signals.

In the case of primates, where the hierarchy in S/N between the two channels is large the mixing of the two channels is significant†. In fact it is not hard to show that the angle of rotation needed is approximately 45 degrees A 45 degree rotation leads finally to the following solutions for the two optimally decorrelated channels with equalized S/N ratios

$$\begin{aligned}
\mathbf{Z}_1 &= (\mathbf{G}+\mathbf{R})\ K_+M_+ + (\mathbf{G}-\mathbf{R})\ K_-M_- \\
&= \mathbf{R}\ (K_+M_+ + K_-M_-) + \mathbf{G}\ (K_+M_+ - K_-M_-) \\
\mathbf{Z}_2 &= -(\mathbf{G}-\mathbf{R})\ K_-M_- + (\mathbf{G}+\mathbf{R})\ K_+M_+ \\
&= \mathbf{R}\ (K_+M_+ - K_-M_-) + \mathbf{G}\ (K_+M_+ + K_-M_-).
\end{aligned} \tag{5.29}$$

Since for primates, $K_+(f)M_+(f)$ and $K_-(f)M_-(f)$ are very different, the end result is a dramatic mixing of space and colour. For example, cell type no. 1 at low frequency has $K_-(f)M_-(f) > K_+(f)M_+(f)$ so it performs an opponent $\mathbf{R}-\mathbf{G}$ processing. As the frequency is increased, however, $K_-(f)M_-(f)$ becomes smaller than $K_+(f)M_+(f)$ and the cell makes a transition to a smoothing $\mathbf{G}+\mathbf{R}$ type processing (Derrington *et al* 1984). In figure 14, we show the filters in frequency space, in the **R** and **G** basis. These filters are in principle directly measurable in contrast sensitivity experiments. We view the zero crossing at some frequency as a generic prediction of this theory.

In figure 15 (dashed line), we show how the solutions look for a typical set of parameters after Fourier transforming back to space. We can see cell type no. 1 summates red mostly from its centre and an opponent green mostly from its surround, while for type 2 the red and green are reversed. These cells can be termed single opponency cells, as seen in primates (Derrington *et al* 1984). One might object that the segregation of the red and green in the centre is not very dramatic. Actually, this is due to the simplified model we have taken. Complete segregation can be achieved if one allows the synaptic noise parameter N_0, which was set to 1 for the dashed line, to be different for the two channels. A difference of 1/2 between the two noises produces the solutions shown by the solid curves in figure 15.

We hope the results of this and the previous section have convinced the reader that the application of information theory to neural systems merits further investigation.

† A rotation could have been done in the goldfish case also, but there the two channels (5.28) $\mathbf{Z}_1$ and $\mathbf{Z}_2$ already have approximately equal S/N so the degree of mixing is very small or ignorable.

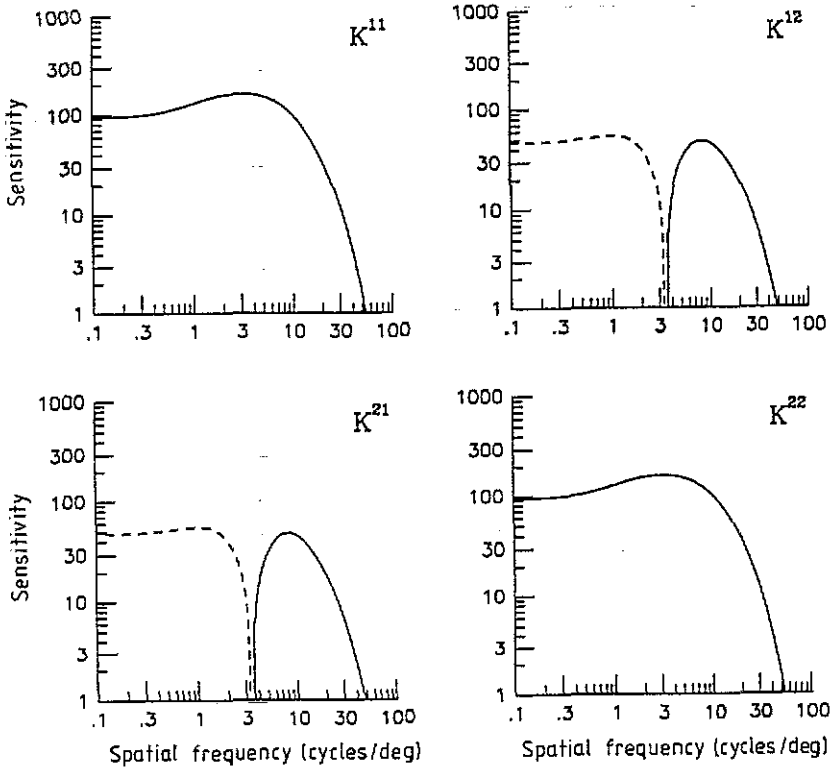

Figure 14. Predicted retinal filter $K^{ab}(f)$ in the **R** and **G** basis for the primate regime $r = 0.85$ and for the same parameters as those in figure 12. The solid (dashed) curves represent excitatory (inhibitory) responses. Notice that both cells $\mathbf{Z}_1$ and $\mathbf{Z}_2$ make a transition at some frequency from opponent colour **G** − **R** or **R** − **G** to non-opponent **G** + **R**.

Acknowledgments

I would like to thank Z Li and N Redlich, L Kruglyak and K Miller for many hours of useful discussions, J Kmiec for comments on the review and L Ferraro for interesting discussions. This work is supported in part by a grant from the Seaver Institute.

Appendix

Another useful concept in information theory is that of mutual information of two events or variables, O and L defined as

$$I(O;L) \equiv H(O) + H(L) - H(O,L)$$

where $H(O,L)$ is the joint information; $H(O,L) = -\sum_{O,L} P(O,L) \log P(O,L)$. This quantity has some interesting properties. For example if the events O and L are completely statistically independent then $P(O,L) = P(O)P(L)$ and $H(O,L) = H(O) + H(L)$ making $I(O;L) = 0$. On the other hand, if the two events are completely dependent then $H(O,L) = H(L) = H(O)$ and $I(O;L)$ is the same

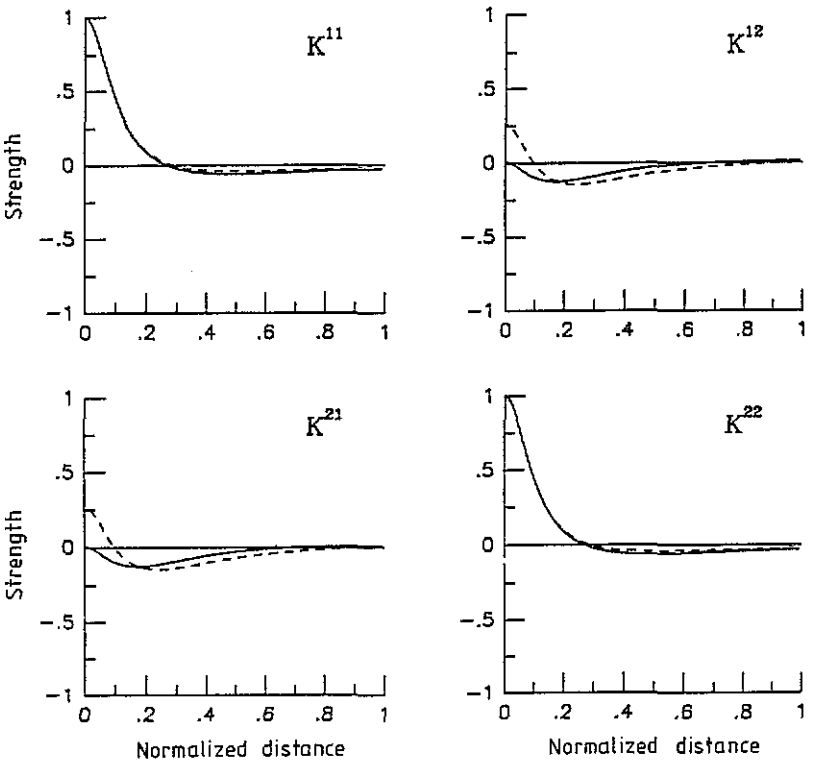

Figure 15. Predicted retinal kernel K^{ab} in the **R** and **G** basis in the primate regime $r = 0.85$ and for the same parameters as those in figure 12 are shown by the dashed curves. The solid curves use the same parameters except that the parameter N_0 was allowed to be different in the luminance and chromatic channels by a factor of two. This was done to illustrate that complete colour segregation in the cell's centre can easily be achieved.

as $H(O)$ (or equivalently $H(L)$). Thus, $I(O;L)$ in general is a measure of the interdependence of the two events. In fact, it can be thought of as the information carried by O about the event L. If $O = L + n$ where n is some additive noise and if all the variables are Gaussian distributed with some variance, then $I(O;L) = H(O)$ – noise entropy, a fact that we needed in our analysis in section 5.2.

References

Atick J J and Redlich A N 1990a Towards a theory of early visual processing *Neural Computation* **2** 308–20

—— 1990b Quantitative tests of a theory of retinal processing: contrast sensitivity curves *Report* Inst. for Advanced Study IASSNS-HEP-90/51, submitted for publication

—— 1992a What does the retina know about natural scenes? *Neural Computation* **4** 196–210

—— 1992b Convergent algorithm for sensory receptive field development *Report* Inst. for Advanced Study IASSNS-HEP-91/80, submitted for publication

Atick J J, Li Z and Redlich A N 1990 Color coding and its interaction with spatiotemporal processing in the retina *Report* Inst. for Advanced Study IASSNS-HEP-90/75

—— 1991 Understanding retinal color coding from first principles *Report* Inst. for Advanced Study IASSNS-HEP-91/1 (1992 *Neural Computation* in press)

—— 1992 What does post-adaptation color appearance reveal about cortical color representation? *Report* Inst. for Advanced Study IASSNS-HEP-92/7, submitted for publication

Attneave F 1954 Some informational aspects of visual perception *Psychol. Rev.* **61** 183–93

Barlow H B 1961 Possiple principles underlying the transformation of sensory messages *Sensory Communication* ed W A Rosenblith (Cambridge, MA: MIT Press)

—— 1985 Perception: what quantitative laws govern the acquisition of knowledge from the senses? *Functions of the Brain* ed C W Coen (Oxford: Clarendon)

—— 1989 Unsupervised learning *Neural Computation* **1** 295–311

Barlow H B, Hawken M, Kaushal T P and Parker A J 1987 Human contrast discrimination and the contrast discrimination of cortical neurons *J. Opt. Soc. Am.* A **4** 2366–71

Barlow H B and Foldiak P 1989 *The Computing Neuron* (Reading, MA: Addison-Wesley)

Barlow H B, Kaushal T P and Mitchison G J 1989 Finding minimum entropy codes *Neural Computation* **1** 412–23

Barnard G A 1955 Statistical calculations of word entropies for four western languages *IRE Trans. Inform. Theory* IT-1 49–53

Bialek W 1990 Theoretical physics meets experimental neurobiology *Lectures in Complex Systems (SFI Studies in the Sciences of Complexity, Lecture vol II)* ed E Jen (Menlo Park, CA: Addison-Wesley)

Bialek W, Rieke F, de Ruyter van Steveninck R R and Warland D 1991 Reading a neural code *Science* **252** 1854–57

Bialek W, Ruderman D L and Zee A 1991 Optimal sampling of natural images: A design principle for the visual system? *Advances in Neural Information Processing Systems* vol 3, ed R P Lippmann, J E Moody and D S Touretzky (San Mateo, CA: Morgan Kaufman)

Buchsbaum G and Gottschalk A 1983 Trichromacy, opponent colours coding and optimum colour information transmission in the retina *Proc. R. Soc. Lond.* B **220** 89–113

Campbell F W and Gubisch R W 1966 Optical quality of the human eye *J. Physiol.* **186** 558–78

Davson H 1980 *Physiology of the Eye* (London: Academic)

Daw N W 1968. Colour-coded ganglion cells in the goldfish retina: extension of their receptive fields by means of new stimuli *J. Physiol.* **197** 567–92

De Monasterio F M, McCrane E P, Newlander J K and Shein S J 1985 Density profiles of blue-sensitive cones along the horizontal meridian of macaque retina *Invest. Opthalmol. Vis. Sci.* **26** 289–302

Derrington A M, Krauskopf J and Lennie P 1984 Chromatic mechanisms in lateral geniculate nucleus of macaque *J. Physiol.* **357** 241–65

De Valois R L, Morgan H and Snodderly D M 1974 Psychophysical studies of monkey vision: III spatial luminance contrast sensitivity tests of macaque and human observers *Vision Res.* **14** 75–81

Enroth-Cugell C and Robson J G 1966 The contrast sensitivity of retinal ganglion cells of the cat *J. Physiol.* **187** 517–52

Field D J 1987. Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* A **4** 2379–94

—— 1989 What the statistics of natural images tell us about visual coding *Human Vision, Visual Processing and Digital Display (Proc. SPIE 1077)* pp 269–76

Gallager R G 1968 *Information Theory and Reliable Communication* (New York: Wiley)

Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6** 721–41

Goodall M C 1960 Performance of stochastic net *Nature* **185** 557–8

Gouriet G G 1952 A method of measuring television picture detail *Electron. Eng.* **24** 308

Harrison G W 1952 Experiments with linear prediction in television *Bell System Tech. J.* **31** 764

Hentschel H G and Barlow H B 1991 Minimum-entropy coding with Hopfield networks *Network* **2** 135–48

Hinton G E and Sejnowski T J 1983 Optimal perceptual inference *Proc. IEEE Conf. Computer Vision and Pattern Recognition (Washington, DC)* (Piscataway, NJ: IEEE) pp 448–53

Jacobson H 1951 The informal capacity of the human eye *Science* **113** 292–3

Kelly D H 1972 Adaptation effects on spatio-temporal sine-wave thresholds *Vision Res.* **12** 89–101

Kersten D 1990 Statistical limits to image understanding *Vision: Coding and Efficiency* ed C Blakemore (Cambridge: Cambridge University Press)

Kretzmer E R 1952 Statistics of television signals *Bell System Tech. J.* **31** 751–63

Kornhuber H H 1973 Neural control of input into long term memory: limbic system and amnestic syndrome in man *Memory and Transfer of Information* ed H P Zippel (New York: Plenum)

Laughlin S B 1981 A simple coding procedure enhances a neuron's information capacity *Z. Naturf.* **36c** 910–2

—— 1987 Form and function in retinal processing *Trends Neurosci.* **10** 478–83
—— 1989 Coding efficiency and design in visual processing *Facets of Vision* ed D G Stavenga and R C Hardie (Berlin: Springer)
Linsker R 1988 Self-organization in a perceptual network *Computer* **21** (March 1988) 105–17
—— 1989a An application of the principle of maximum information preservation to linear systems *Advances in Neural Information Processing Systems* vol 1, ed D S Touretzky (San Mateo, CA: Morgan Kaufman) pp 186–94
—— 1989b How to generate ordered maps by maximizing the mutual information between input and output signals *Neural Computation* **1** 402–11
Lythgoe J N 1979 *The Ecology of Vision* (Oxford: Oxford University Press)
Mullen K T 1985 The contrast sensitivity of human colour vision to red–green and blue–yellow chromatic gratings *J. Physiol.* **359** 381–400
Papoulis A 1984 *Probability, Random Variables, and Stochastic Processes* (New York: McGraw-Hill)
Pearlmutter B A and Hinton G E 1986 G-maximization: an unsupervised learning procedure for discovering regularities *Neural Networks for Computing (AIP Conf. Proc. 151)* ed J S Denker (New York: AIP)
Pratt F 1942 *Secret and Urgent* (New York: Blue Ribbon Books)
Redlich A N 1991 Redundancy reduction as a strategy for unsupervised learning *Report* Inst. for Advanced Study IASSNS-HEP-91/87
Schreiber W F 1956 The measurement of third order probability distributions of television signals *IRE Trans. Inform. Theory* **IT-2** 94–105
Shannon C E 1951 Prediction and entropy of printed English *Bell System Tech. J.* **30** 50–64
Shannon C E and Weaver W 1949 *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press)
Shapley R and Enroth-Cugell C 1984. Visual adaptation and retinal gain controls *Prog. Retinal Research* **3** 263–346
Shaw S R 1984 Early visual processing in insects *J. Exp. Biol.* **112** 225–51
Srinivisan M V, Laughlin S B and Dubs A 1982 Predictive coding: a fresh view of inhibition in the retina *Proc. R. Soc. Lond.* B **216** 427–59
Sterling P 1990 Retina *The Synaptic Organization of the Brain* ed G Shepherd (Oxford: Oxford University Press)
Sziklai G 1956 Some studies in the speed of visual perception *IRE Trans. Inform. Theory* **IT-2** 125–8
Uttley A M 1979 *Information Transmission in the Nervous System* (London: Academic)
Van Essen D C and Anderson C H 1988 Information processing strategies and pathways in the primate retina and visual cortex *Introduction to Neural and Electronic Networks* (Orlando, FL: Academic)
Van Essen D C, Olshausen B, Anderson C H and Gallant J L 1991 Pattern recognition, attention, and information bottlenecks in the primate visual system *Conf. on Visual Information Processing: From Neurons to Chips (SPIE Proc. 1473)*
Van Nes F L and Bouman M A 1967 Spatial modultion transfer in the human eye *J. Opt. Soc. Am.* **57** 401–6
Warland D, Landolfa M A, Miller J P and Bialek W 1992 Reading between the spikes in the cercel filiform hair receptors of the cricket *Analysis and Modeling of Neural Sytems* ed F Eeckman (Norwell, MA: Kluwer)
Watanabe S 1981 Pattern recognition as a quest for minimum entropy *Pattern Recognition* **13** 381–7
Watanabe S 1985 *Pattern Recognition: Human and Mechanical* (New York: Wiley)