Visual statistics using neural networks

Roland Baddeley

June 27, 1994

Acknowledgments

This work has been carried out while funded by a SERC grant. Thanks to my supervisor, Roger Watt, for providing most exceptional equipment, for useful comments on the early chapters, and for financial help at the end. A special thanks to Peter Hancock, without him much of the thesis would have been very different, taken much longer, and would not have been nearly as much fun. Though we only directly collaborated on the first two chapters, he provided much useful comment on the rest of the ideas in this thesis, and gracefully informed me when I was talking rubbish.

Thanks also to Ben Craven, for his variable but informed enthusiasm, for help with my writing, and especially for help with the ideas in chapter 5. Many of the ideas in this chapter are directly due to collaboration with Ben, and many less plausible ideas removed after useful discussion. Again Ben helped make it fun.

Thanks also to my second supervisor, Bill Philips, for his extreme enthusiasm, for his ability to see the woods for the trees, for great help in my writing, and for arranging a visit a conference in Spain. Also to Leslie Smith, for financial help when I needed it, and for the ability to have read and usefully commented upon a chapter practically before you handed it to him. To Jim Kay, who could explain addition without being condescending, to Ian Paterson, who if you had a problem, would help by providing 5 references, 10 possible solutions, and have probably have half of them implemented by the morning. Thanks also to Steve Dakin (sorry about frying tuna), Daniel Sturdy for inspired conversations at 4.00am in the morning, Steve Emott for the swift halves, Will Goodall-Lelong for the application of the plastic stick, Dario and Paul for cigarette breaks, to Tony for showing me how to turbo charge a Morris, Todd for his mean fields, to David Cairns for help with the work on oscillators and pointing out the spelling mistake in the last sentence. And also to all the other members of the CCCN.

Thanks and love to my Mum, Dad, and especially Fiona.

Abstract

This thesis describes the application of statistical techniques to natural images as a means of gaining insight into the operation of low level vision. First, the statistical technique of principal component analysis is applied to a collection of natural images: a match with psychophysical data is found; and a solution to the dynamic range problem proposed. The problem of learning and calibrating psychological and physiological representations of space is then investigated. The grey level correlations in natural images are measured and their physical causes investigated. The resulting correlations are related both to psychological distortions of space and to the cortical representation of space in V1 in macaque monkey. The interpretation in terms of a system calibrating itself using the correlations in the input signals is shown to produce accurate psychological and physiological predictions. Lastly the problems of creating low level models of the visual input is looked at using a framework originally proposed by Hinton and Sejnowski (1983). The way in which phase coherence of (neuronal) firing in a network can label the probability of an interpretation is demonstrated. A new search technique, inspired by the different time courses of inhibition and excitation in the cortex, is proposed for searching for the most likely visual interpretation. It is concluded that statistical techniques can provide insight into the operation of low level vision.

Contents

Chapter 1

Introduction

SUMMARY This thesis examines low-level biological vision using a statistical approach. It attempts to gain insight into three separate low-level processes using this framework: the extraction of "features", the organisation of the spatial relationships between these features, and the capture of a statistical model of the features. These three processes are loosely identified with the creation of receptive fields, the maintainance of topographic organisation, and the action of the horizontal connections within the cortex. These processes and their relevance to low-level vision are outlined in the first half of this chapter.

One common and strong argument against approaching low-level vision from a statistical point of view, is that it ignores the behavioral significance of objects. For some amphibians, the visual system is highly specialised to recognise the objects it needs to react to in order to survive: for example flies. If this were true of primates, then a statistical approach that treated the whole of the visual signal as important would be misguided. The second half of this chapter argues that for an adaptive and general-purpose visual system, there are a number of reasons why low-level vision could act in ignorance of behavioural significance. The question then becomes an empirical one: does treating low-level vision as a statistical process produce insight into its operation? The rest of this thesis attempts to answer this question for the three problems of the extraction of representations, the maintenance of spatial relations, and the maintaining of a statistical model of the relationship between elements in the representation.

1.1 What is low-level vision?

In many psychological theories of perception, between the outside world and "object recognition", lies an intermediate stage of processing; low-level vision. A simple definition is that this is the stage of visual processing where the elements of the representation are not "objects", but lower level constituents. Examples of such proposed representational elements are lines, edges, or zero crossings. For many years it has been known that certain neocortical areas are involved in this visual computation. This may or may not provide additional constraints on the form of this computation, but allows an alternative working definition: the area of interest in this thesis is the form of computation potentially carried out in or before areas V2 and V1 in the cortex, the first two neocortical visual areas (Zeki, 1978).

What forms of computation is being performed at this level? Of these, for which could insight be gained from a statistical perspective? This thesis will concentrate on three forms of computation that are probabaly general across the modalities of perception, important to perception, and able to be related to both to psychological/ psychophysical and physiological phenomena.

1.1.1 Receptive field/feature extraction

The most striking aspect of the computation in V1/V2 is that it is understandable. Thirty years after Hubel and Wiesel's first reports (Hubel & Wiesel, 1959), it seems clear that the neurons in early visual cortex can be, at least roughly, characterised in terms of their receptive fields. This is highly understandable in contrast to potential and previously proposed forms of computation. Even if this simple framework misses many of the details, it does capture something of the operation of these early visual neurons.

What is less clear is why the receptive fields are of this form. Why orientation sensitivity? Why spatial differentials? There is no lack of hypotheses: for example (Marr, 1982) proposed that the neurons are acting as edge/zero crossing detectors. (Koenderink & van Doorn, 1987) that they are for approximating the image using a local Taylor series approximation , (Campbell & Robson, 1966) that these cells perform a local Fourier analysis, or that both spatial and frequency based attributes are important and should be optimally represented (Daugman, 1984). In this thesis yet another hypothesis is proposed, that receptive fields are the way they are because they are tuned to maximise the amount of information conveyed about the visual world. Although this

proposal may not be incompatible with previous proposals, it has the strong advantage that it is testable. The statistics of visual images can be estimated, the corresponding "optimal" filters calculated, and these filters can be compared to those found both psychophysically and physiologically. This is described in chapters 2, and 3.

1.1.2 The topographic organisation of visual representations

Another striking aspect of the representation occurring in the early visual areas is its topographic organisation. Theoretically, there could be no relationship between the physical location of a neuron, and the location in the visual world represented by that neuron. In mammalian vision this is definitely not the case. Neurons that represent nearby features in the world are physically close in the cortex: the representation and the world is in topographic correspondence.

There are a number of questions that can be asked about the map-like character of the cortex. How is it set up? How is the map kept in correspondence with the world in the face of changing optics? What kind of signals are used in forming the map, and what determines the final form of the map? There has been much neurophysiological work on these maps, and a number of neural network algorithms have been proposed. This thesis examines whether knowledge of the visual input statistics, a potential source of calibration signals, can provide insight into the structure of these cortical maps.

A related problem is that of keeping our psychological representation of space in registration with the world. If the eye changes shape or is damaged, if someone puts on glasses, or if there is neural damage, then the mapping between the world and our representation of the world changes. Again the statistics of the world could provide a set of calibration signals much as the test card allows the calibration of television sets. Chapter 4, investigates the possibility that at least part of the calibration of psychological space is achieved using the statistics of naturally encountered images.

1.1.3 Modelling the relationship between features

The last aspect of potential computation in V1/V2 that will be described is the creation of models of the relationship between features. Orientation information is integrated from "line detectors" more efficiently than if the line detectors were treated as independent (Andrews, 1967a), and Gestalt psychologists have created a number of demonstrations where the percept is understandable only if there are interactions between the parts. Cells in V2 respond to "illusory contours", lines that have been inferred to be

present, although there is only limited evidence for this inference occurring in in V1 (see for instance (von der Heydt & Baumgartner, 1984)).

It is worth considering whether many of these perceptual completion phenomena can be captured via the application of simple rules, such as: if in previous experience the presence of two collinear line elements increases the probability of a third, then if I am sure of the presence of two such line elements, my estimate of the probability of a third should be increased. This kind of inference can be treated as the application of a probabilistic model. The conditional probabilities of various feature combinations can be estimated from images, and a network can be used to apply this knowledge to the estimated scene characteristics. The horizontal connections in cortex provide a potential substrate for implementing such a model. These connections are of the form required: strong connectivity between similar orientations. The responses of V1/V2 cells show context sensitivity, as does human psychophysical performance. Inferences of this type are (thus) a potentially important aspect of low-level vision.

If this behaviour is seen in a statistical inference framework, then a number of different questions can be asked. What kind of networks can perform such inference? Given the vast number of possible combinations of features, how can we find the most probable? Does the neurophysiology provide any clues as to how this search is performed? This thesis looks at how one aspect of the neurophysiology, the slow time-course of one of the inhibition systems, can be used to improve such a search in a way that produces reasonable answers throughout the search, in contrast to previous techniques that produce unusable answers during most of the search.

When applying previous knowledge to a visual input, identifying the most probable state of the world is the primary interest. Also of interest is the probability of this interpretation relative to others. An interpretation made at very low light levels for instance may be the most probable but still not be very good. If we have sources of more reliable information, then the lack of confidence must be signalled. Chapter 6 argues that one way that the cortex could signal both what the most probable interpretation, and how probable it is, is to code the probability in terms of the phase coherence of neuronal spiking. This means of signalling is then tested for plausibility in a simplified network model of spiking neurons. It is argued that this gives a way of interpreting the observed coherent activity in cortex.

1.2 Is a behaviourally ignorant approach relevant?

Given that the three processes (finding of "features", maintaining the topograpic organisation, and modeling the relationships within a representation), are potentially important components of low-level vision, and that a statistical perspective is applicable, there is an important theoretical consideration that needs clearing up. This thesis treats visual experience as unlabelled data, and applies statistical methods to this data in the hope that this will shed light on the processing occurring in early mammalian vision. The relationship between this process and the tasks of vision, recognising a friend or guiding a walk along a rocky path for instance, is sometimes obscure. What is required of vision is a system that allows the manipulation of the world, and the identification of behaviourally relevant inputs. Can an approach that ignores the behavioural importance of the world be valid?

Vision is surely for finding behaviourally relevant information in the world. For a fly that can only land on something, eat and avoid being eaten, representing all the characteristics of the visual world is of little use. The requirements of vision in the fly are different from those of image reconstruction or inverse optics as studied in machine vision. In image reconstruction, all the input is treated as important and the task is to use knowledge to infer the most likely image. In the fly, the only aspects of the visual world that need to be represented are those that make a difference to its behaviour. Exactly recovering the colour of an object is pointless if the only difference that it makes to a fly is whether the object is green or not. Surely, then, statistics is only a useful approach if we know which elements of the signal are important? For humans this is a daunting task.

In this thesis, representations are determined by the statistical form of the input: little attention is paid to its behavioural relevance. For studying the fly this approach is surely wrong. Nothing in the signal specifies that the optic flow is the important aspect of the signal- this only becomes clear with knowledge of the behaviour of the fly. If low-level vision in higher anamals is to be treated as being largely determined by the structure of the input, then this has to be justified for reasons other than the pragmatic one of having little quantitative knowledge of what is behaviourally relevant.

One can argue that representations in higher anamals can be treated as predominantly determined by the signal because these creatures have much richer behaviour. A fly has a very limited repertoire for interacting with the world, so its representation only has to deal with few behaviourally relevant differences. The world models of

man, monkeys, and cats are much richer and many more visual world differences are behaviourally relevant. This implies that ignoring behavioural relevance can possibly be justified. At a low-level, all variations in the signal are of potential interest, and therefore should be represented. The richer the models used by the behaving agent, the better this approximation holds, and for humans and monkeys, with behaviour that takes into account much richer aspects of the signal than the fly, the approximation that any regularity and any structure in the signal is important, is more reasonable. This argument, however, seems unsatisfactory and for more convincing support we have to look to learning.

1.2.1 Why adaptation is important in low-level vision

Learning of concepts, and words, is always recognised as important by cognitive science. Within visual science, and especially vision studied psychophysically, the visual system is often modeled as pre-determined and static. On viewing an object, the system performs a pre–determined set of operations on the input, and communicates a useful symbolic representation to higher and more flexible levels of cognition. Although this may be a useful fiction when studying higher aspects of vision, this is neither good engineering for a system, nor consistent with observed behaviour.

Even the simple visual system of the *Limulus* displays adaptation to the environment: continued stimulation decreases the response of the eye to stimulation. Variation in the visual diet of cats changes its visual representations (Blakemore & Cooper, 1970). A human with amblyopia, even if corrected, will not have the information to learn the exact structure of a stereo representation. When the visual field alters due to age related changes in the eye, the wearing of glasses, or the simple shifting of rods and cones in the eye, the system has to learn the new meanings of the measurements.

1.2.2 Simple learning

If a snail (eg *Helix albolabris*) is placed on a wooden platform, which is jerked back and forth, the snail retreats into its shell. If this procedure is repeated a number of times, slowly the snail's reaction becomes smaller and smaller (Carthy, 1958). This is not spectacular learning: never-the-less it is useful. Note how this simple process can separate out behaviourally relevant stimuli: the snail learns that even large variations in the world can be ignored if they suffer no harmful effects. Continued use of this method will selectively remove representations of variations that are not behaviourally relevant.

This type of learning can be extended to more complicated things. Samuel designed a draughts playing program that learned to improve its performance and eventually to beat him (Samuel, 1967). The program's "vision" consisted of a number of position evaluation systems; it had to learn which of these was useful in finding out if it was winning: i.e, it had to use experience to create a representation of what was behaviourally relevant. This problem is a *credit assignment* problem: if the program wins, which of the evaluation systems should be rewarded. This problem was solved by rewarding all the experts that had been used a lot when it won, and penalising the experts used when it lost. This form of learning, known now as reinforcement learning (Barto *et al.*, 1983; Sutton & Barto, 1991) was sufficient to allow the system to learn to play a reasonable game of draughts. It would also be good enough to allow a bee to learn different shapes with a reward of food, or a worm to navigate a maze.

1.2.3 The credit assignment problem

Reinforcement learning is very useful, and will allow the crafting of representations that directly mirror the behaviourally relevant aspects of the world. The trouble is that such a system does not scale well. For example, if we had two layers of experts, one looking at the draught board and one looking at the results of these experts, the credit assignment problem becomes much harder. How do we know how to reinforce the lower layer experts? This problem held back neural network research for many years until two methods, Boltzmann machines and back-propagation, provided at least partial answers. By allowing the approximate calculation of the contribution of the low-level experts to the behaviour; credit, and therefore a training signal, could be given to networks with a number of different levels.

For systems with a highly accurate description of the errors, this provides a method for creating intermediate representations that are tailored to the required behaviour. Even when there are a number of intermediate representations between the rewards and the input, the form of the intermediate representations can be calculated, and this allows the simple of networks with one intermediate level of representation. Two intermediate levels can also be learnt with more time and more accurate measures of the errors. With three intermediate levels, huge amounts of computation sometimes allows the generation of appropriate lower-level representations. For greater numbers of intermediate levels between the input and the source of reinforement (reward or punishment), learning determined purely by errors is plagued with great difficulties, and the human visual system is clearly of this sort. In the human visual system, the signal passes through

at least four representations (LGN, V1, V2, V4) before it reaches areas with access to the objects' identity, and therefore reward and punishment. Even ignoring the possibly inaccurate representation of the errors, the credit assignment problem is very great if we want direct behavioural relevance to determine the intermediate representations. The error signals after being passed down through the intermediate levels are too diffuse and the learning signal to the low levels in a system with noise becomes effectively random.

This problem is not just a quirk of back-propagation or Boltzmann machines. If changes in a companies profitability were used by a company clerk to determine if he kept his records on paper or on computer, then because so many other factors will effect the companies profitability, his decision will effectively be random. Many factors affect profitability and the effect of this one change would be drowned out. V1 is going to receive little reinforcement from behavioural relevance if deciding whether to use long or short "line detectors". The further from the representations that their behavioural relevance can be assessed, the worse this problem is. Although genetics can provide intermediate representations that are approximately correct, if any adaptive behaviour is required of these low-level representations, the credit assignment problem means that the behavioural relevance signal will be very diffuse and practically useless. Either we abandon adaptation in all but the highest levels of representation, or we find an alternative learning signal to supplement behavioural reinforcement in early visual processing.

1.3 Simplifying complex problems: informational accountancy

The company clerk can make decisions on the form of his filing system, and the visual system can calibrate its retinotopic representation of space and the form of its receptive fields. If a signal based purely on behavioural relevance is too noisy and unstable to be useful at this low-level, some other signal must be used. Learning in networks that generate representations without a reinforcement signal is known as unsupervised learning. Networks of this type form representations based on internally generated rules, rather than those determined by the requirements of a particular task. Inherent in these rules is some notion of what is a good representation: rules of informational housekeeping. These allow the system to create a representation with (good) characteristics defined with respect to a particular input.

One action a clerk can perform on information is to summarise, to order and structure. It may be that this information, when passed up to higher levels is never used, but if it is needed, then expression in a succinct form, ordered, and with irrelevancies removed, will make its use much easier. Statistical information can be used to represent in many fewer variables a signal that has redundancies. For a higher level, with better knowledge of what features of the signal are important, a representation that abstracts the data and expresses it succinctly will be useful.

The system can also make guesses as to which measurements are important. If the government finds that industrial output is down, employment is down, and the value of the pound is down, then it is reasonable to suppose that there is some deep underlying cause for all these things and should be noted. Again, if something is detected using touch, sound, and vision, then the measurements probably refer not to some random fluctuation in the environment but some underlying cause, and this again can be used to modify representations.

Another useful operation is to give measurements a context. If a local manager is told that a workshop used 400 resistors last week, this means little. If also told that on an average week only 10 are used, then he can tell that something is probably wrong. Statistical history can provide an intrinsic scale for measurements, in its simplest form by expressing measurements in terms of standard deviations from the mean.

The inputs can also be sorted for relevance. If no other information that the clerk was using changed whenever a red truck drove past, then this should not be reported. If every time this happened 390 resistors went missing, then it should be. Even without detailed knowledge, if some measurements are known to be important, then this can be used to guide the weight given to other measurements based on their statistical relationship to the ones of known importance. This concept will be applied to the formation of cortical maps where it is assumed that the system is interested in what it is fixated on. This is then used as an organisational principle to arrange the geometry of measurements from the periphery.

1.3.1 Forming a model of the input

After working at a job for a long time, a clerk will build up a model of all the measurements he works with. Although he may not know the significance of what he is recording, as he is exposed to large amounts of data, he becomes much more likely to notice if one of the measurements has a mistake in it and therefore be able to correct it. If a particular file went missing, then asking the clerk the probable number of resistors

used is more likely to get an accurate answer than asking the area manager.

Model forming is also proposed to have a role in low-level vision. In a noisy environment the best place to guess about the presence or absence of a line is at a low-level with its access both to vast experience of lines and to more information about other lines currently in view. Even with no knowledge of the importance of various measurements, low-level vision can form models of these measurements, provide guesses when not all information is present, and label inputs as like or unlike what has been seen before.

1.4 The statistical approach

There are many ways of studying the operation of the visual system. Even if low-level vision partly uses statistical processes to develop its representations, why should this be of interest? Many other ways have been found to study vision, which do not require assumptions about the lack of a behavioural relevance signal. The representation develops in a certain way; we can find out what this final representation is, and we can study how it is used. How does viewing low-level vision as a statistical process help?

This thesis does not report any tests on the representations which would demonstrate whether or not a particular form of representation helped the learning of visual tasks. However the data is at least similar to that encountered in the visual world. Much of computer vision uses a single picture of a model (Lena), a book of textures (the Brodatz book), or collections of crosses and T's as representatives of visual experience. These may well be good "toy world" problems containing many of the characteristics of real world images, but previous work in A.I. has found that many solutions to toy problems are not robust.

Computer vision, viewed as a way of inferring the structure of the world from a static image, proposes that in order to use such information as stereo, shading, and texture, assumptions about the form of surfaces need to be made: e.g.,the surfaces have locally bounded differentials, or can be approximated as low order polynomials (Poggio *et al.*, 1985; Horn & Schunck, 1981). This can lead to impressive computer models for performing inverse optics, but inherent within these models is a regulariser of how the world usually is. The assumptions may or may not be good, but testing on synthetic images, or a few real ones, will not tell you how good they are.

By keeping the models very simple, parallel, and fast, it is possible to expose them to large numbers of real images. This means that they can complement more complicated

models for which the time to analyse a single image runs into hours. Even if they contain extreme simplifications (e.g., the universal use of correlation as a measure of relatedness), exposure to large numbers of images can reveal information that could not be found from the more limited number of images that can be studied with more complete and complicated models. The statistical models proposed are all very simple. The analysis of 1,000 images can regularly be performed in the time many models take to analyse one.

Another reason that models of this form are worth studying is the proposed ignorance about low-level representations of behavioural significance. This makes them inherently testable. A model that proposes very good performance is difficult to test. A model that proposes that many of the characteristics of low-level vision are generated through ignorance will make mistakes. If a book store shelves quantum field theories under gardening, and gauge theories under metal work, it can be inferred that the sorting is not being performed by someone with knowledge of the books' contents. If a visual system constantly overestimates some distances, and underestimates others, it can be inferred that the system does not "know" about distance, but is inferring it from some other characteristic. If some characteristic is found that gives the same distortions, then this is evidence that the system might not be using distance, but this other characteristic.

1.5 The structure of this thesis

The question "are the representations for low-level vision being moulded by statistical constraints" is an empirical one. The hypothesis might be wrong for two reasons, the low-level representations used could be entirely genetically predetermined, or the brain might have a good solution to the multiple-level credit assignment problem. This would render inferences made after ignoring the behavioural relevance of the inputs very suspect. The rest of this thesis consists of a number of investigations to determine whether this approach produces reasonable and testable results. It is split into three sections, as follows.

Chapters 2 and 3 considers statistical feature extraction. After discussing some of the desirable characteristics of representations, one particular form of statistical analysis is investigated: principal component analysis (PCA). Although this has some limitations, it is mathematically well understood and has many of the characteristics desired of a good representational technique. It also has the virtue of having a fast, simple neural

network implementation which allows large numbers of natural images to be analysed. The results are compared to measured human performance in judging orientation, and the problems of variable illumination are investigated.

Chapters 4, and 5, investigates how the psychological and physiological correlates of distance can be approximately learned. If the structure of cortical maps is determined by the need to represent the point of fixation, and the periphery is represented in proportion to its measured relevance to this fixation point, then correlation can be used to define the metric. After sampling from a large collection of images, the geometry of the resulting representation is inferred on the assumption that correlation is used as the metric. This is compared quantitatively to empirical measurements on the geometry of V1 in the macaque monkey.

Finally, problems associated with forming and applying statistical models in low-level vision are addressed. The problem of signalling how probable a possibly noisy input is to higher levels is looked at, and the proposed solution is compared to the neurophysiological evidence. Then the problem of performing the massively parallel search required to fit a model to the data, and the requirement of communicating useful intermediate interpretations are addressed. The solution to these problems, inspired by biology, is compared to earlier computational techniques and psychophysical results.

Chapter 2

The Principal Components of Natural Images

SUMMARY

The principal components of 40 real world images are found using a neural network technique. The initial components are a Gaussian followed by directional, first derivative operators strongly aligned to the vertical and horizontal, then moving to successively higher order operators. The use of these principal components both for representation and for insight into the statistics of scenes is discussed. Two of the operators resemble oriented bar detectors and the orientation tuning curves match very closely those found psychophysically by Foster and Ward (1991). This is shown to be consistent with the orientation sensitivities being determined by a measured vertical-horizontal anisotropy in the image statistics[1].

2.1 Introduction

Principal component analysis (PCA) is a technique that has found widespread use in signal processing, data compression and statistics. It is a dimension reduction technique where rather than use all N variables in a data set X, we use a smaller number d (where $d < N$) of new variables that are formed as standardised[2] linear combinations of the original variables. The first principle component (PC) is the standardised linear combination of the variables that has the largest variance. The second PC is the standardised linear combination of the data variables that has the largest variance

[1]This work was carried out in collaboration with Peter Hancock and was published in two papers (Baddeley & Hancock, 1991; Hancock *et al.*, 1992).

[2]The weightings of all variables are normalised to length one.

whilst being uncorrelated with the first component, and the k'th component is the the standardised combination that has the largest variance whilst being uncorrelated with the previous k-1 components. If we wish to reduce the number of variables used to represent an image, but wished to capture as much of the variance in the image as possible, then the linear combinations of variables we should chose should be the principle components.

Mathematically, this problem of finding linear projections that maximise the variance accounted for, is identified with an eigen-vector problem. Specifically the principle components are the eigen-vectors of the covariance matrix of the image measurements, and the order of the principle components is determined by the order of the eigen-values associated with the eigen-vectors. If

$$\mathbf{Q} = [Covar(x_i, x_j)] = [< (x_i - \bar{x_i})(x_j - \bar{x_j}) >] \tag{2.1}$$

with x_i being the i'th element of x, $< ... >$ being the average value across samples, and $\bar{x_i}$ being the average value of x_i, then the principal components (PCs) of a distribution are defined as the solutions $\mathbf{a}$ to the equations :

$$(\mathbf{Q} - \lambda \mathbf{I})\mathbf{a} = 0 \tag{2.2}$$

again with the weighting matrixes standardised to unit length

$$||\mathbf{a}|| = 1 \tag{2.3}$$

here, the λ's are the eigenvalues, and $\mathbf{a}$ are the principal components. These principal components (PCs) can then be ordered in terms of decreasing λ.

Principal components can provide a new representation of an image. Rather than representing an image in terms of the original grey level measurements, we can describe an image in terms of its projection onto the PCs. As a new representation, this has a number of virtues, as follows,

- If we measure 4000 different image locations, it is unlikely that we are measuring 4000 independent processes. A reasonable heuristic (Occams razor) is therefore to describe as much variation in the image in as succinct a form as possible. It is hoped that a description in terms of these fewer variables will better capture the structure of the underlying image generation processes. Unfortuetly this is purely a heuristic and there is no garantee that that the reduced representation will make the underlying structure any clearer.

- The information (conditional entropy) of a Gaussian variable with independent identically distributed gaussian noise added is a simple monotonicaly increasing function of the signal to output noise ratio. Therefore, in this Gaussian case, by maximising the variance of projections, we also maximise the transmited information. Therefore, given a limitted number of variables n, and Gaussian assumptions, the principle components (and rotations of them) are the linear combinations of the variables that maximise transmitted information.

- The PC's are orthogonal and uncorrelated. The lack of correlation between the variables is important because the speed of many network computations is determined by the off-diagonal elements in the correlation matrix. Examples include the operation of surface interpolation networks (Pentland, 1993), and supervised neural networks (Foldiak, 1992).

- Given data that is clustered, it may be desirable to use vaiables that capture between cluster differences. If we assume that between-cluster variation is greater than within-cluster variation, finding directions of high variance is also likely to produce directions that allow the discrimination between clusters. Therefore PCA that finds projections of large variation will, heuristically, also capture projections that distinguish between clusters (Friedman, 1987).

- PCA can also be seen as a noise-removal technique. If the noise on all variables is independent but the signal displays higher order correlations, capturing projections of consistent variation will suppress noise and emphasise signal.

- If looking for higher order statistical regularities, these will be confounded by the simple second order statistical structure. By expressing the data in terms of the PCs, all first and second order structure is removed from the data. This process, known as sphering the data, potentially makes the finding of higher order structure simpler (e.g., (Friedman, 1987)).

- The distance (scaled by the standard deviation) of a sample from the average sample is not a good way of describing how different this sample is from the mean; it confounds likely variations with unlikely ones. By transforming to the axis given by PCA (and dividing by each PCs standard deviation), this becomes a more sensible measure, the Mahalanobis distance (Duda & Hart, 1973).

- PCA is also interesting because it has a number of simple, parallel, neural network

implementations (Oja, 1982; Sanger, 1989a; Sanger, 1989b; Leen, 1991). These systems have two parts: simple Hebbian learning (Hebb, 1949), and a mechanism for ensuring the orthogonality of the components. The simplicity of these systems means that biologically possible Hebbian implementations exist (Foldiak, 1992).

PCA has drawbacks as well. In natural images most of the variance is in the low spatial frequencies. PCA by attempting to account for the variance, emphises the low spatial frequencies at the expense of the higher spatial frequencies.

Bossomaier and Snyder (Bossomaier & Snyder, 1986) suggest that the response properties of simple cells are similar to those generated by PCA and give this as a possible explanation for the spatial frequency organisation of the cortex . Daugmann (1990) gives as a virtue of the Gabor representation[3] its resemblance to PCA. However both these claims are based on theoretical considerations rather than the analysis of natural images. Plumbley (1991) found Information-theoretic, optimal, linear representations and suggests neural network implementations. In the case of equal Gaussian noise on all inputs, as assumed here for simplicity, this reduces to PCA.

Field (1987) analysed the statistics of six images of rocks and trees using Fourier techniques, revealing a $1/f^2$ power spectrum. He suggests that PCA will suppress high frequency information and therefore is inappropriate for image representation. An alternative log Gabor[4] representation is proposed that equally represents low and high frequency. Atick et al (Atick *et al.*, 1990; Atick & Redlich, 1992) used the statistics found by Field, the chromatic correlations induced by the spectrally overlapping red and green receptors, together with information-theoretic measures, to derive optimal spatial and chromatic tuning curves. The results match well with the observed human contrast sensitivity function, and the chromatic organisation of ganglion cells in goldfish and primates.

Linsker (Linsker, 1986) presents a multiple layer model of the development of cells in the visual cortex. Given initially random input, he shows that both center surround and

[3]Gabor considered the problem of measuring both the spatial location and frequency of a one dimensional signal (Gabor, 1946). He showed that if we wished to measure both simultaneously, then we came across a fundamental uncertainty. Accurate knowledge of the position implied an inaccurate estimate of the frequency and vice versa. The filter that minimises this uncertainty in position and frequency is a sine (or cosine) multiplied by a Gaussian envelope and this is known by his name. Daugman (Daugman, 1985) extended the concept to two dimensions, and this has been influential, mainly because the two dimensional Gabor resembles receptive fields in V1.

[4]The Gabor filter is a Gaussian in the frequency domain. Because the spectral power in natural scenes as measured by Field was constant in log frequency space, Field proposed that a filter that was gaussian in log frequency space would be more appropriate for image analysis. This filter he named the log-Gabor.

oriented bar-detector cells developed as found in cortex. This result was subsequently shown (MacKay & Miller, 1990a) to be interpretable as the result of the interactions of the PCs of the input statistics.

2.2 Experiment 1: the principal components of natural images

2.2.1 Method

Forty images were obtained by scanning photographs at a resolution of 300 dots per inch and 256 grey levels. These pictures consisted of 15 natural scenes (Figure 2.1), 15 taken in and around our laboratory (Figure 2.2) and ten assorted pictures from other sources. No attempt was made to correct for any optical irregularities in the lenses, since a variety of cameras were used for the original photographs. Each image was reduced to 256 pixels square. Sixty four by sixty four pixel samples were chosen by selecting an image and an area within it at random. This relatively large size was chosen to reduce any effects of pixellation. The population mean grey level, as estimated from 20,000 samples, was subtracted from each pixel value because the network requires zero mean input to converge.

If the PCs are taken of rectangle sections of the image, the distance from the centre to the corners is further than from the center to the edges. Because the correlations in images decay with distance, and these correlations determine the PCs, this will distort the PCs aligning them to the geometry of the sampling rectangle. The effects of rectangular sampling geometry were avoided by windowing the data, much as is done when using a Hamming window in Fourier analysis. This windowing has been justified by Linsker in terms of "the probability of connection", decaying with distance. Here it is justified because we are interested in the statistics of the images rather than the edge effects or the details of sampling. Windowing the data lessens the effects of the geometry of the sample on the results.

Specifically, the sample was windowed using a Gaussian with a standard deviation of ten pixels. This means that the borders of the square sample are more than three standard deviations from the center: sufficient to avoid edge effects. The sample was then normalised to make the sum of the squares of the pixel values equal to one.

The first PC of a data set can be obtained using a simple Hebbian neural network algorithm (proposed by (Oja, 1982)). To extract further PCs, we used a generalisation of this rule (Sanger, 1989a). This was done in preference to using conventional matrix

Figure 2.1: The 15 natural images used in Experiment 1.

diagonalisation techniques to examine the computational characteristics of the neural network technique. Proofs exist both of convergence and stability of the final results of Sanger's network (Leen, 1991). Despite this, there were potential problems associated with a large scale implementation (e.g., convergence speed, stability with a finite learning set, the required precision of parameters, and problems with the particular data set used). If the brain operates in a manner similar to the proposed algorithm, then these would also be problems that it would face. By using the neural network, any such deficiencies might be revealed. When I wrote the code I believed also that this method would be both faster and more memory efficient for the very high dimensional inputs used here. There are claims to this effect in the original Sanger Thesis. This has proved not to be true. Current methods for extracting eigenvectors are highly developed and if only the first few are required, very fast and memory efficient.

Oja (1982) shows that if a single unit performed a weighted sum of its inputs, and updates its weights with both a Hebbian term and a weight decay term ($-\gamma w_i$) as follows

$$y = \sum_{i=1}^{N} x_i w_i \tag{2.4}$$

$$\Delta w_i = \eta y(x_i - y w_i) \tag{2.5}$$

Figure 2.2: The 15 scenes taken in and around the office.

then given a long enough exposure to the input statistics the weight vector of a single unit will converge to the direction of the first PC (the Hebbian term), and be of length one (the weight decay term).

This rule was generalised (Sanger, 1989a) to multiple units by replacing the update rule in equation 2.4 by

$$\Delta w_{ij} = \eta y_j (x_i - \sum_{k=1}^{j} y_k w_{ik})$$

With this method, a good approximation to the first 15 PCs was found in tractable time[5]. The weights were initially set by drawing from a uniform random distribution, centred on zero, and having a small width to break symmetry.

Training proceeded by first choosing a random image from the image set, then choosing a random 64×64 sample from within that image. After preprocessing (described above), this sample was used as an input to a network consisting of 15 units. The weights were updated accordingly. This process was repeated 120,000 times at which point, the weight vectors were stable. Convergence is assisted by 'annealing' the learning rate: typically it began at 1.0 and was halved every 20,000 presentations.

2.2.2 Results

The results of the network's computations on a particular image set are 15 PCs. These can be visualised by plotting the weights as a 64X64 grey level image, the weight size

[5]For a more detailed description of why this will converge, see (Hertz *et al.*, 1990). For detailed proof include the necessary stability calculations, see (Leen, 1991)

being represented by grey level. The 15 PCs extracted from the entire 40 image set are shown in Figure 2.3. The first component is a slightly anisotropic blob and is well approximated as a Gaussian. The least squares fitting Gaussian has a standard deviation 9.1 in the vertical direction, and 9.8 in the horizontal. That this component is not determined entirely by the windowing function but also by the statistics is shown both by this anisotropy and by other studies on pages of text where the first component is no longer even approximated by a Gaussian (Hancock *et al.*, 1992).

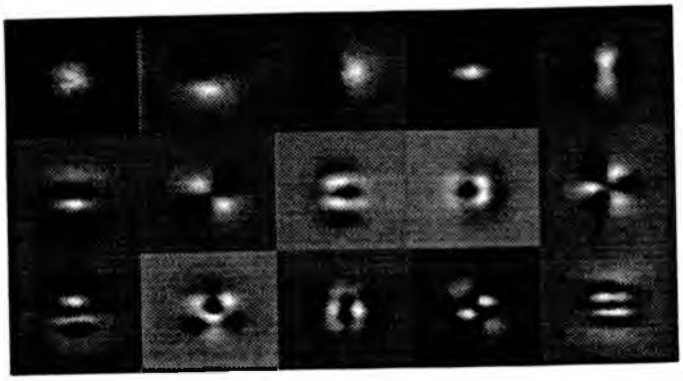

Figure 2.3: Experiment 1:The first 15 principal components of our images.

Components two and three resemble directional derivatives of a Gaussian. The order of the components is not random; the vertical operator was always first, followed by the horizontal operator. It was confirmed that this was not an artefact of the digitising process by rescanning the images oriented at 45°. The orientation of the operators moved around with the images as shown for the first six components in Figure 2.4. It was also confirmed that the components were not caused by the particular scale tested by performing the same analysis, this time on samples of 128×128 and 32×32. The gross structure of the components was unaffected.

As a test for sensitivity to the particular image set chosen, the subset of images shown in 2.1 was tested and again the results were similar.

The fourth and fifth components resemble oriented "bar-detectors". Based on psychophysical experiments on the detectability of oriented line segments in a background of other line segments, Foster and Ward (1991) also proposed the existence of two oriented "bar-detectors". These two detectors centred on the vertical and horizontal have different orientation tuning curves. The orientation tuning curves of the two bar detectors produced by PCA were measured and compared to those found by (Foster &

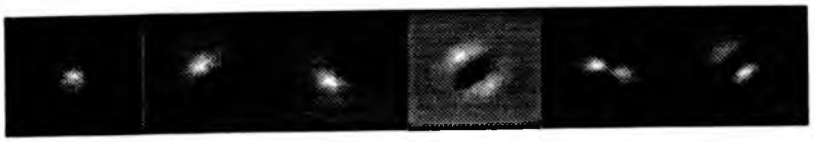

Figure 2.4: Experiment 1:The first 6 PCs of 15 images, when rotated by 45 degrees on the scanner.

Ward, 1991). As can be seen in Figure 2.5, the match is very good.

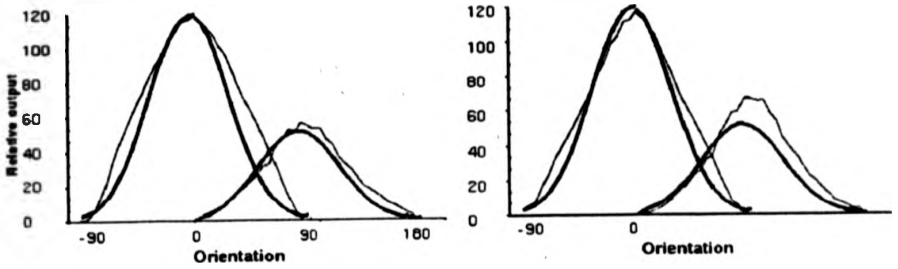

Figure 2.5: The orientation sensitivity model of Foster and Ward (1991) (thick lines) with "bar-detector" components from (a) the subset of 15 images (thin lines) and (b) the full image set. Vertical units are arbitrary.

2.3 Discussion

2.3.1 The Anisotropy in the orientation tuning curves and the visual statistics

The PCs are determined by the eigenvectors of the correlation matrix: the statistical structure of the natural images is reflected in the structure of the PCs found. The first principal component has greater extent in the horizontal direction than the vertical: the second PC for both image sets was firmly aligned as a vertical operator. This indicates that there is an anisotropy of the statistics of images in the vertical relative to the horizontal direction, i.e., the correlations decay more slowly in the horizontal direction than the vertical.

To investigate whether this anisotropy was the reason for the different orientation tuning curves of the vertical and horizontal bar detectors, the following was tried. A number of fractals were generated which were Brownian in the luminance domain.

We used fractals because they have information at all scales and the correlations will be approximately similar to those encountered across the ensemble of natural images. Therefore we could see what effect a vertical horizontal anisotropy in these correlations has on the PCs, independent of the effects of the exact spatial frequency spectrum of natural images.

These fractals were generated by dropping edges on an image at random angles; the statistics can be made anisotropic by biasing the orientation of the edges. This was achieved by drawing the edges from a rectangular distribution, allowing the biasing of lines in the horizontal direction. The vertical to horizontal ratio of the distribution was varied between 0.9 and 0.3 in steps of 0.1. At each of these seven levels of anisotropy, the first seven PCs were calculated and the orientation tuning curves found. The results are shown in Figures 2.6 and 2.7.

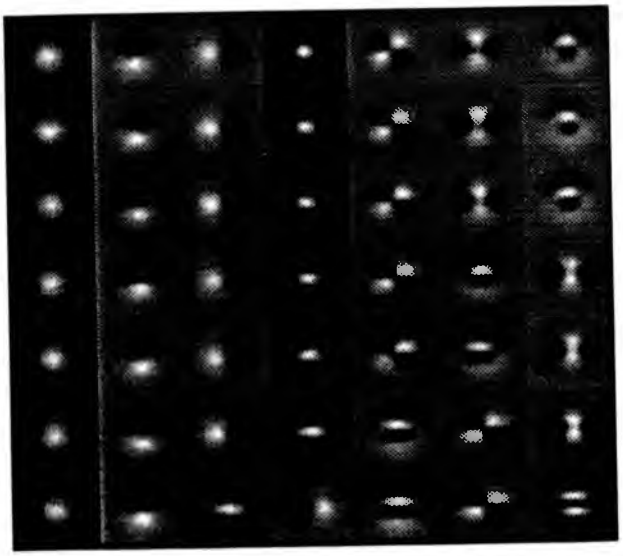

Figure 2.6: First seven PCs from Brownian fractals, varying the vertical to horizontal ratio from 0.9 (top) to 0.3 in steps of 0.1.

As can be seen, the ratio of the orientation tuning curves varies systematically with the anisotropy of the images. The ratio that produces the observed anisotropy in the orientation tuning curves is about 0.55. An analysis of the correlation structure of the images using techniques outlined in Chapter 4 shows that there is an anisotropy of the decay in correlation in the vertical to horizontal direction of about 0.6. This leads to the interesting hypothesis that the observed anisotropy found psychophysically is

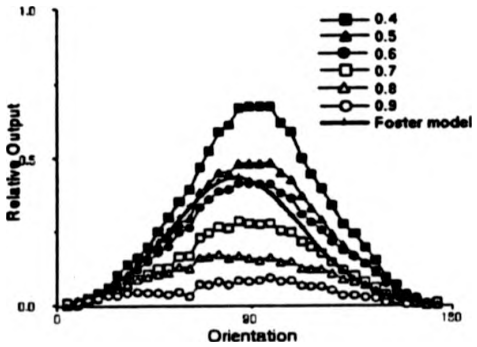

Figure 2.7: Orientation sensitivity of vertical 'bar-detectors', varying the anisotropy of Brownian fractals, adjusted such that the matching horizontal component has a peak output of one. Foster and Ward's model is included for comparison (thick line).

caused by the representations used by the brain tuning themselves to the statistics of natural images.

A second interpretation is that it is not the orientation tuning curves that are dependent on the image statistics but the image geometry itself. By changing this geometry, the orientation tuning of units will also change. The implications of this have not been investigated.

2.3.2 The orthogonality constraint

The receptive fields generated by PCA, and those inferred from short exposure duration psychophysics, are orthogonal, whilst those of receptive fields measured in cortex are certainly not orthogonal. Daugman (1988) addresses this problem from the opposite direction where he proposes a non-orthogonal Gabor representation. This better resembles the orientation tuning curves of cells with long duration simulation, but in its simplest form does not possess the advantages of orthogonality.

To gain the advantages of orthogonality Daugman proposed a neural network. Given a non orthogonal set of Gabor filters and an image, this network finds the coefficients that represent the image so that projecting the coefficients back gives the best reconstruction of the image in terms of least mean squared error. These coefficients have advantages over the simple multiplication of the filter with the image. They possess some of the advantages of orthogonality, and the code exploits the redundancy in the image. An over-complete basis set can produce sparse codes which are easier to work

with in neural networks, and the image is represented by the nearest matched filter, which gives signal-to-noise advantages over representations using orthogonal filters[6].

Unfortunately finding these coefficients takes time for the gradient descent on the reconstruction, and during this search the coefficients are not correct. Only filters that are orthogonal will give the correct coefficients in the first pass through the system and this suggests a possible alternative method for finding the correct coefficients. The filters initially used are both broad and orthogonal, allowing instant discovery of the correct coefficients. The filters then decrease their orientation bandwidth and intermediate filters are used. The rough orthogonal solution is found, and this is used as a guide to finding the non orthogonal solution.

This suggestion is at least consistent with neurophysiology: in cats the cells that fire on the presentation of an input show an excess of horizontal and vertical detectors: 84 out of 116 cells are within 20° of horizontal or vertical. While cells firing later show no such anisotropy (Vidyasagar & Henry, 1990)). In 50% of cells in V1 the orientation tuning curves start broad and then become progressively more tightly tuned. Within some of these cells the preferred orientation also changes over time (Dinse & Best, 1990). The initial broad orientation bandwidth and preference for the horizontal and vertical directions could be what was being measured by Foster and Ward since they used very short presentation times.

2.4 Conclusion

In conclusion, the PCs have a number of advantages as a representation of a signal. That psychophysics on very short exposure duration stimuli is well modeled by two orthogonal filters with the same orientation tuning curves as produced by PCA is intriguing. For longer exposure times, both psychophysically and neurophysiologically, orientation tuning curves are much tighter and non–orthogonal (Daugman, 1990).

This opens up an interesting possibility. Watt (1987) has suggested that the visual

[6]The matched filter is a useful concept from linear filtering theory. It is relevant when we are trying to detect the presence of a signal in the presence of noise. What is required is a filter that is maximally sensitive to the signal, whilst being minimally sensitive to the noise and this can be quantified as the signal to noise ratio. Under certain conditions on the noise, and only allowed linear filters, the optimal filter (with the highest signal to noise ratio), is known as the matched filter. The impulse response (the shape of the weights) of this filter resembles the object that is searched for. Put more simply, the matched filter concept states that if we wish to detect a sine wave in a signal, the optimal filter is a sine wave shaped one, and not a square wave (or any other form). Here the searched for signal is not known, but by representing this signal by the nearest filter, we gain the signal to noise advantages of matched filtering.

system scans from course to fine spatial scales over time . This scanning, it is argued, makes optimization simpler; it may be the case that orientation space is scanned in addition to scale space. This could speed up the finding of the correct coefficients in a network such as Daugman's. If we were to scan in orientation space, the PCs would be a good starting point, they are orthogonal allowing the coefficients to be found purely by finding the dot product with the image, and PCs contain the most information for a limited number of coefficients. It is not suggested that PC type receptive fields operate in the cortex, but that such an analysis offers a good approximation to the system when starting with the initial rough orientation specificity.

Chapter 3

Illumination and the importance of angular information

Summary

In natural scenes, the strength of the illumination can change over a range of 15 orders of magnitude. This means that the same scene, with probably the same behavioral relevance, can generate images that vary greatly. Changing the illumination will change the absolute values of measurements, but it will not change the differences between the logarithms of image intensity measurements. Therefore a logarithmic transform is proposed as a part of the preprocessing stage to eliminate problems of input dynamic range. The PCs of logarithmically transformed images are found and shown to be very similar to those resulting from straight PCA.

The preprocessing of the images by taking logs, then normalising, is similar to the preprocessing performed in a form of statistical analysis. This analysis, known as log contrast principal component analysis (LCPCA), is also performed on images and again produces very similar components. The orientation tuning curves are also similar to those found in the previous chapter indicating that the form of the components is not the result of the particular non linearities or preprocessing used, but a robust result of the image statistics. A theoretical framework from Mackay and Miller (1990) produces similar components allowing the empirical components to be classified within a framework. Doing this shows that with the anisotropic statistics of the natural images, increasing orientation resolution accounts for more image variation, than increasing spatial scale resolution. The human visual system also shows relatively greater resolution for orientation than scale.

3.1 Introduction

PCA was described in chapter 2. This chapter proposes that because of the nature of the visual world, there are very large multiples of the measured brightness (caused by changes in the illumination) that have little or no relevance to the identification of objects in the world. This leads to the proposal of an alternative form of statistical analysis based on the contrast (ratios of measured brightness), rather than the measures themselves. This simply translates to performing our processing upon the normalised logarithm (log) of the measured grey levels, rather than on the measured grey levels themselves.

The problems of using the covariance (the measure of relatedness often used in PCA) are first described, and the rest of the chapter explores the additional observations that can be made if our analysis is based upon the measured image contrasts, rather than on simply the measured luminace.

3.2 Problems with covariance as a measure of correlation

At a given light level I_1, the interrelatedness of two pixels (a and b) as measured by the covariance (the measure sometimes used in PCA) is proportional to:

$$Covar(I_1 R_a \cos\theta_a, I_1 R_b \cos\theta_b) \tag{3.1}$$

Since lighting is constant, this measured the average correlation of two measurements of $R\cos(\theta)$, an interesting measurement that relates to the spatial structure of the world measured from.

Problems start occurring when we want to improve these statistical estimates by taking more measurements. Now the illumination is I_2. Although the world is the same, our new measurements are now based on a different illumination and the covariance is then multiplied by a factor of $(I_2/I_1)^2$. If we average over the two samplings, we does not improve our estimate of the relationship within the ensemble of scenes, but get a more complicated measurement, dominated by the brighter lighting condition, with the lesser lighting condition acting purely as noise to our estimates. If there are many lighting conditions, as is common even within a scene, then the problem becomes acute. Although we can make good estimates of the correlation of the incoming light, these are dominated by the effects of illumination, which varies by a many orders of

magnitude, and has little relevance to spatial correlations in the world that generated them.

To circumvent this problem, we need a measure of correlation that is sensitive to differences in the world, but is constant under different illuminants. Since changes in illumination correspond to different multiplication of the measurements, we need a measure of correlation that is invariant over the multiplication of all values within a sample. If two pixels are close spatially, and we assume that they receive the same illumination within each sample, then since the illuminant is constant, the ratio of the values will be invariant over different lighting conditions.

One way obtain such a measure is by finding the covariance of the ratios but if there is a large numbers of measurements, N, this requires the storing of $N^2/2$ ratios. Alternatively we can convert ratio relationships to difference relationships by taking logs. This means that the differences between values are now representative, but each has a constant added to it, proportional to the log of the illumination:

$$\log(IR_a \cos\theta_a) = \log(R_a \cos\theta_a) + \log(I) \tag{3.2}$$

To convert differences to absolute values, we just express them relative to the mean of the log values within a sample:

$$\mathbf{x}' = \log(\mathbf{x}/g(\mathbf{x})) \tag{3.3}$$

equivalently

$$x'_i = \log(x_i) - mean(\log(\mathbf{x})) \tag{3.4}$$

where $\mathbf{x}'$ are the transformed variables, $\mathbf{mean(x)}$ is the average within the sample and $g(\mathbf{x})$ is the geometric mean of the values in a sample $\mathbf{x}$. This produces samples that are invariant with respect to multiplication and hence invariant over changes in illumination.

Converting the values to the new form and calculating the covariance of these measures can be thought of as using an alternative measure of covariance that is sensitive to ratios rather than differences. This measure of correlation is known as the centred log-ratio covariance $\mathbf{\Gamma}$ (Aitchison, 1986). This will be sensitive to regularities of reflectance structure in the world, not just the illumination created by this structure.

More formally if :-

$$\Gamma_{i,j} = Covar(\log(x_i) - mean(\log(\mathbf{x})), \log(x_j) - mean(\log(\mathbf{x}))) \tag{3.5}$$

where $\mathbf{x}$ is the vector of measurements from one sample, x_n is the n'th sample, and $mean(\log(\mathbf{x}))$ is the average log value within that sample. $\Gamma_{i,j}$ is then the centred log ratio covariance.

If we replace the covariances $\mathbf{Q}$, with the centered log ratio covariances Γ:

$$(\Gamma - \lambda \mathbf{I})\mathbf{a} = 0 \tag{3.6}$$

the solutions for a correspond to those of log contrast principal components (LCPCs) (Aitchison, 1986), the ratio equivalent to PCA. For the visual system, trying to analyse regularities in the reflectance, not in the incoming light, this is a more appropriate analysis.

3.3 Experiment 2: Log contrast PCs of natural images

A neural network was used to find the log contrast principal components of natural images. Thirty images were randomly chosen from the 40 images used in chapter 2. These consisted of a number of natural outdoor and indoor images, each of size 256 by 256 pixels, each normalised so that the grey level are between 1 and 256. Details of the images can be found in chapter 2.

Sanger's (Sanger, 1989a) version of Oja's (Oja, 1982) principal component net was used. When exposed to mean zero data this creates a set of units corresponding to the principal components of the inputs. The data for this net was preprocessed so that the net was performing LCPCA in the following manner:

- A randomly chosen 64 × 64 subsection was extracted from an image randomly selected from the 30.
- The natural log of all the intensity values within a sample was found.
- The within-sample average log value was calculated and subtracted from every value.
- To remove edge effects, the input data was then windowed with a Gaussian of standard deviation ten pixels.

The network was run for 80,000 iterations at which point the projections ceased to change.

3.3.1 Results and Discussion

The results of this computation are shown in Figure 3.1. For comparison the original PCs are also shown in Figure 3.2. Note that the polarity of the components is not meaningful.

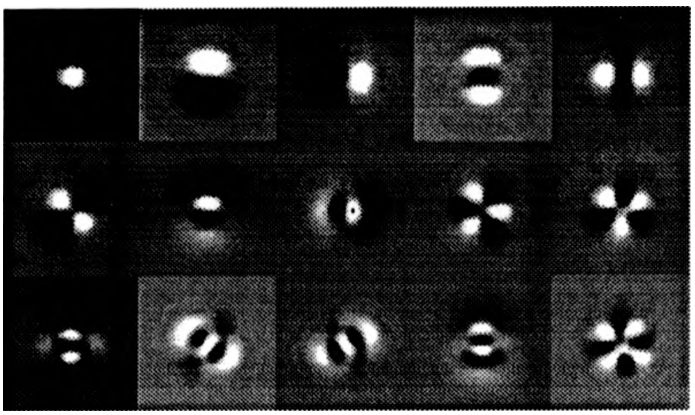

Figure 3.1: The first 15 log contrast principal components extracted from 30 natural images.

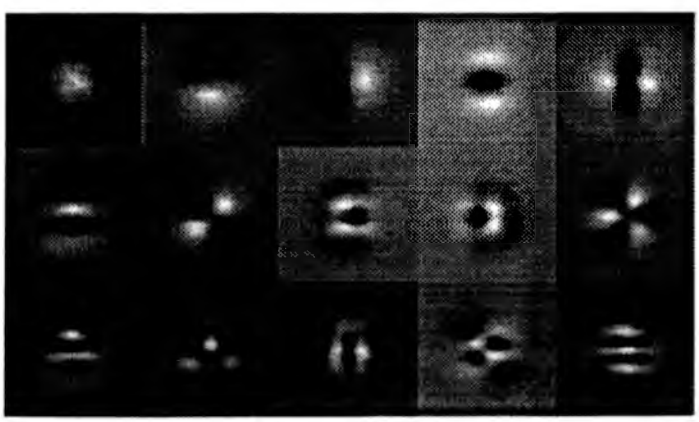

Figure 3.2: The first 15 principal components (with each sample normalised to variance one) extracted from 40 natural images. These are the same as in (Baddeley and Hancock 1991) and are presented for comparison.

Bossomaier and Snyder (1986) propose that one of the strategies used by the visual system is redundancy reduction. This lead them to propose PCA as an optimal rep-

Component number	variance σ^2	number of radial nodes	number of angular nodes	total number of nodes
1	30.0798	0	0	0
2	17.0276	0	1	1
3	11.9143	0	1	1
4	6.3392	0+1	2+0	1+ 2
5	5.7139	0+1	2+0	1+2
6	5.1811	0	2	2
7	4.6955	1	1	2
8	4.1240	1	1	2
9	3.1196	0	3	3
10	3.0486	0	3	3
11	2.3747	2	1	3
12	2.3332	2	1	3
13	2.4151	2	1	3
14	2.0135	3	1	4
15	1.9856	0	4	4
Sum:	102.3658	12	22	34
σ weighted node sum		20.1	43.3	

Figure 3.3: The 15 log contrast PCs. The components are numbered in terms of descending eigenvalue. The estimated average variance found when convolved with the log–contrast normalised samples is also shown. This is proportional to the component's eigenvalue. The number of angular and radial modes is a description of the operator in terms of the quantum mechanical scheme as proposed in (Mackay and Miller,1990) (see text and figure). With components 4 and 5 the plus sign refers to the fact that these two components are formed out of mixtures of two components from MacKay's scheme, one of structure 0,2 and one of 1,0. The σ weighted sum refers to a sum over the angular and radial modes weighted by the standard deviation (proportional to the square root of the eigen value), of each component. Note that both the ratio of angular to radial nodes, and the ratio weighted by standard deviation is very roughly 2:1.

resentation, and a local Fourier expansion to be a good approximation to this. Whilst the principal components of a translation invariant one-dimensional process are sines and cosines of different frequencies, the two dimensional projections as found here and produced by a sampling technique that ensures that the input statistics are translation invariant, are definitely not all spatially localised Fourier basis functions.

Daugman (1990) also suggests that the principal components will be approximated by a localised Fourier analysis and uses this as an argument for the Gabor representation scheme. Although some of the projections seem to be spatially localised spatial frequency operators (Gabors), the "cartwheel" like projections (6,9,10, 15 in Figure 3.1

and 8,9,10,12,13,14, and 15 in Figure 3.2) are not, having a more complicated structure incompatible with a Gabor type interpretation.

3.3.2 The principal components of isotropic Gaussian correlated images

The insight into the presence of the problem components can be found from some analytical work on Linsker's (Linsker, 1986) simulations of receptive field growth performed by (MacKay & Miller, 1990b). Although not specifically interested in PCA, there results depended on finding the eigen-vectors (and hence principle components) of samples from "images". These "images" were not real images, but were analytically defined two dimensional arrays where the correlation between any two points in the image depended only on a Gaussian function of the distance between the points[1]. As in this study, samples from these "images" were windowed using a Gaussian. For this system, it is possible to analytically calculate the form if not the exact ordering of the PC's and these are shown in Figure 3.4.

The first PC is always a Gaussian, the next component is then constrained to be orthogonal. This can be achieved in two ways. Firstly the next component can radially symmetrically increase its spatial frequency resolution, the next orthogonal component of this form is the second differential of a Gaussian (see Figure 3.4). Alternatively it can increase its resolution in the orientation domain. For the Gaussian correlated case studied, this was of the form $r\cos(\theta)\exp^{-r^2/R}$ where θ and r are the polar coordinates of the receptive field. Each step again can expand a previous one by increasing the orientation resolution (increased angular nodes: cart wheel detectors), or increased isotropic spatial frequency (increased radial nodes). Figure 3.4 shows the possible expansions. The eigenvalue of any component is then constrained to be lower than any components higher in the tree.

The PCAs and LCPCAs relative to the analytic components

Although the rigid constraint of rotational invariance is not met for natural images [2], the PCs approximately fit into this scheme. Possibly because of the noise introduced by the variance normalisation of the samples, the later components appear as admixtures of the analytic components. The LCPCs fit very neatly into the scheme (the number of

[1]Note therefore that the correlation is the same in all directions: it is isotropic.

[2]The correlations are flattened due to foreshortening, and show pronounced structure in the vertical and horizontal directions. See later.

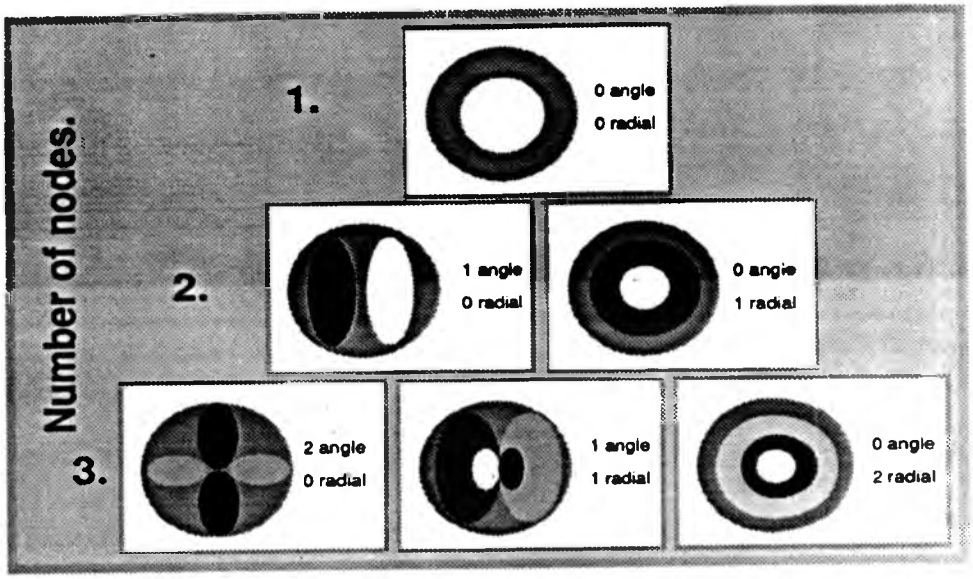

Figure 3.4: The pyramid of PCs of a translationally and rotationally invariant correlation matrix. Because the correlation matrix has this symmetry, so do the resulting PCs. The "Gaussian" receptive field is always the first principal component. One of the two receptive fields directly below it has the second highest eigenvalue. Generally a receptive field type will always have an eigenvalue lower (and hence be a higher order PC) than all receptive fields above it in the tree $(3,1 > 2,1 > 1,1)$ but not necessarily all receptive fields above it $(2,0 <> 0,1)$. In MacKay's analysis the eigenvalues are similar for each level decreasing by a multiplicative constant as each level in the tree is descended. Although the conditions are not met for an analytic derivation of the PCs of natural images, these components provide a framework for understanding the empirically derived components.

angular and radial nodes for each component is given in table 3.3). The only exceptions are components 4 and 5 which are mixtures of a radial and an angular node component.

The constraints of rotational and translational invariance only specify that a component will have an eigenvalue lower than the ones above it in the tree. In the images studied, the amount of variance accounted for was not equal for radial expansions and angular expansions. With n components and generating component $n+1$, this component is constrained to be orthogonal. If the system is similar to the Gaussian correlated case studied by MacKay and Miller, this can be achieved either by increasing the angular resolution, or by increasing the radial resolution. In the images studied, the system

always expanded angularly before it expanded radially. Qualitatively this indicates that angular information is more informative (in that it allows for more of the signal to be accounted for).

If the goal of a representation is to capture as much of the variation in a portion of an image as possible, then if the image statistics are similar to those of the images measured here, orientation (angular nodes) information is more important to represent than spatial frequency (radial nodes).

It is possible to approximately quantify this difference in a number of ways. For the 15 components calculated, the sum of the angular and radial node number of the component weighted by the eigenvalue (variance of the component) can be found. This gives a value of 52 for the angular nodes, 35 for the radial nodes, and a ratio of ≈ 1.5. The sum of the angular and radial nodes can be calculated for the first 15 nodes: this gives a ratio of 1.8. The ratio of the node numbers weighted by the standard deviation can be calculated, this gives a ratio of 2.1 (see Table 3.3). Without a theory of how the information from different channels is combined, the meaning of these ratios is unclear. Despite this, all measures give preference to the orientation expansion and ratio of this to spatial scale of roughly 1.5–2.0. Note this is not a characteristic of PCA but a characteristic of the statistics of the world.

The importance of angular information provides an explanation for the "cartwheel" detectors. Since orientation variation accounts for more variation, but the projections are constrained to be orthogonal, the receptive fields form cart wheel detectors. In the Gabor representation scheme, where the constraint is not maximising information and hence the receptive fields can be non orthogonal, this constraint can be met by having multiple non orthogonal oriented projections. Unfortunately the author knows no technique for creating optimal non-orthogonal receptive fields relative to some statistical environment[3].

The importance of orientation resolution as opposed to spatial frequency information has also been found both physiologically: Jones and Palmer (1987) found a spectral trade off of 2:1 for angular frequency against spatial, Movshon (1979) found a ratio of 1.7:1, and psychologically: Daugman (1984) again found a ratio 2:1. These Figures relate to the aspect ratio of the filter response in the Fourier domain and are not directly related to the what is found here. Despite this, that when LCPCA when is exposed to natural images gives preference to orientation and the difference is in the correct range,

[3]Two possible methods are those of exploratory projection pursuit which has been investigated by the author, and the technique of noise dependent information maximisation developed by Linsker (Linsker, 1992)

may begin to provide an insight as to why the brain also displays additional angular resolution.

3.4 Experiment 3: an image transform approximation to LCPCA

The sample preprocessing inherent in LCPCA (the taking of logs then the "centering" each sample to mean zero), is shown to be well approximated as an image transform. The image transform LCPCAs of a different set of natural images are then found. An image transform equivalent of the variance normalisation used in previous studies is also derived allowing a comparison of the pre-processed images.

A different set of 25 randomly chosen images were used. These consisted of natural scenes such as mountains, animals in natural environments, and vegetation. These images were again size 256x256 and the grey levels were normalised to lie between 1 and 256.

The preprocessing used in LCPCA is sample based. This has limitations both for implementation (requiring separate preprocessing for each sample), and for visualising the effects of the preprocessing. Taking the logarithm of the sample can be replaced by a global log transform of the image. The "centering" or normalising of each sample to mean zero requires an approximation.

Normalising each sample to mean zero rests on the assumption that the entire sample has the same illuminant. Alternatively we can estimate the weighted local mean at every location and remove it. This leaves the problem of what do we mean by local? Ideally we want to include in our measurements all pixels that share the same illuminant. Unfortunately this area would be different each time. Another alternative would be to use a Gaussian weighting function to estimate the local mean. This would be appropriate if there was one scale (σ) that illumination changes took place over, but in a world that can be viewed from many different distances, this is not the case. Ideally we want to treat nearby and far images in the same manner, we would want a filter that has no intrinsic scale. Instead of a Gaussian weighting function, the mean was estimated via an exponential of time constant $\alpha = 10$.

This created the following process for transforming the images:

$$M = D \times \log I \tag{3.7}$$

$$I'_{ij} = \log(I_{ij}) - M_{ij} \tag{3.8}$$

Where $I_{i,j}$ is the original image intensity at i,j, $I'_{i,j}$ is the transformed intensity, $M_{i,j}$ is the locally weighted mean log at point i,j, and × indicates convolution [4]. The local weighting function D has the form:

$$D(x,y) = \frac{1}{2\pi\alpha^2} e^{-\frac{1}{\alpha}\sqrt{x^2+y^2}} \tag{3.9}$$

This preprocessing results in log transformed images that are locally (in terms of D) mean zero. This process was applied to all 25 images, the results of this were then processed by a Sanger network. The only preprocessing used in the Sanger net was a Gaussian windowing of the input data to remove edge effects. A sample size of 32x32 was used to speed convergence.

In Chapter 2, the problem of variable sources of illumination was addressed by normalising each sample to mean zero and unit variance. This has been widely used both implicitly (expressing filter outputs in terms of standard deviations from the mean is effectively normalising the whole image for variance for example (Watt, 1991)), and explicitly (see for instance (Caelli & Moraglia, 1986)).

To allow a direct comparison to the LCPCA preprocessing, an image transform equivalent of locally normalising to mean zero, variance 1 was created:

First the image was transformed to locally mean 0 by subtracting the local mean as estimated via a Gaussian:

$$I'(x,y) = I_(x,y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(i^2+j^2)}{\sigma^2}\right) I(x+i, y+j) didj \tag{3.10}$$

Then by normalising all values to local unit variance.

$$I''(x,y) = I'(x,y) \Big/ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(i^2+j^2)}{\sigma^2}\right) I'(x+i, y+j)^2 didj \tag{3.11}$$

The results of performing straight PCA on the log contrast pre-processed images is shown in Figure 3.5.

The two forms of normalisation

To compare the results of the two image transformation techniques, the two techniques were applied to two images with very different illumination within the picture. The results are shown in Figure 3.6.

[4] The images were assumed to have reflected copies attached at the edges for the purposes of convolution. This results in superior results than toroidal boundary conditions.

Figure 3.5: The image transform log contrast PCs based on 25 images.

3.4.1 The relationship of "real" and image transform log contrast PCs

Although the image based preprocessing will be equivalent to a different high pass transform of the (logarithm of) the image, the resulting PCs are near identical. The first 11 are all equivalent (save for some distortions possibly introduced by the smaller image set), the differences in the last four components are small and are possibly due to differences in the two image sets or to sampling variability.

The relation to biology of the image transform

Whereas the sample based normalisation used in the straight LCPCA has no obvious biological interpretation, the image based normalisation process - $I'_{i,j} = \log(I_{i,j}) - M(x,y)$ now has a very simple interpretation: an on-centre off-surround filter operating on the log of the image values. The retinal receptors are known to have an approximately log transform for a large part of their range[5]. The proposal that the lateral inhibition is used to remove the amount of activity that can be predicted from the illuminant is similar to the proposal of that the role of the inhibition in the retina is to remove predictable variation from the signal (Srinivasan *et al.*, 1982). Since much of the predictable variation will be caused by common illuminant, the resulting inhibitory region will be similar and (Srinivasan *et al.*, 1982) provide evidence that the center surround ganglion cells in the fly are well modeled by such an assumption.

[5]The approximately log transform region of response occurs around the level of illumination the eye is currently adapted to (Baylor & Fuortes, 1970; Normann & Perelman, 1979).

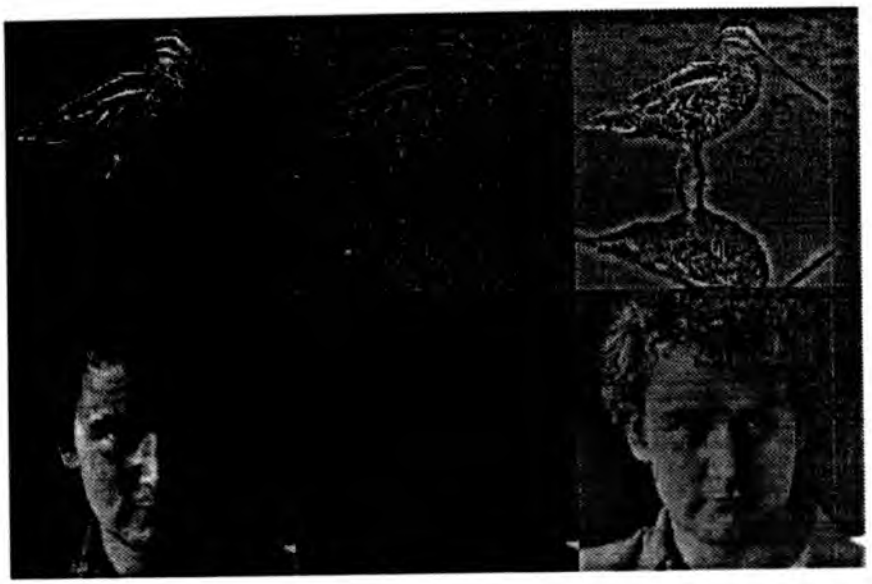

Figure 3.6: The result of applying the two preprocessing techniques to two images with very variable illumination within the images. The images to the left show the original images, the two center images show the effect of normalising the images to local mean zero/variance 1, the two images to the right show the effect of performing the log–contrast preprocessing.

Log contrast preprocessing versus using the raw or variance normalised images

Figure 3.6 shows the results of performing the two kinds of normalisation on the images. Both images have two similar parts containing images under different illuminations: the two sides of the head, and the bird and its reflection. Both sides of the head contain the same information and it would be desirable if both sides were treated similarly.

The first two images are the raw unprocessed images. Although both sides of the head, and the bird and its reflection have similar structure, they have very different illumination and hence brightness. For a system based on simple intensity, such as PCA, the bright parts of the image will be treated differently from the dark. The resulting components will be dominated by the extreme intensities.

The second two pictures show the result of variance normalisation using two different scales. Both do a good job of making all parts of the image equally relevant for the calculation of the correlation structure. In both, the brightness of both halves of

the face, and the bird and its reflection are very similar. Unfortunately as well as normalising this structure, additional structure is introduced. Areas where there is a lot of visual structure (such as the bird's body, around the eye brows) have had this structure suppressed. More importantly, in areas of almost uniform brightness such as the sand around the bird, the prescription that all areas are of equal variance turns these areas into noise. In the previous studies, since many of the images contained such uniform regions (for example sections of sky), the PCs found will be distorted by the effects of the preprocessing induced noise. This may explain why the resulting components did not fit into angular/radial classification scheme as well as the results of the LCPCA. Local variance normalisation rests on the assumption that all areas of the image are of equal interest. When this is not so, as in areas of uniform intensity such as the sky, by forcing them to have equal variance, random fluctuations are enhanced increasing the noise.

The last two pictures show the effects of applying the log–contrast preprocessing. Although in both images there is evidence of ringing of the filters [6], the removal of the illumination effects is good (both sides of the face are similar, as is the bird and its reflection). In the regions of low illumination (the birds reflection, the right side of the face), the image and therefore the statistics have been normalised to be the same as the bright equivalent. More importantly, in the regions of little grey level structure, false structure has not been introduced.

The comparison of "neural images" is not meant to imply a picture in the head hypothesis but is used purely as a means of comparing the form of information made explicit using different forms of preprocessing. Using raw images creates problems of gain control, only the brightest and darkest parts of the image contribute significantly to estimates of the correlation. Variance normalisation overcomes this problem but at the expense of suppressing structure in "interesting" regions and amplifying noise in unstructured regions (the same argument can be applied to whole image variance normalisation). Although log contrast preprocessing is not perfect, it appears to introduce less artifacts than the other two techniques and performs well at removing the confounding effects of variable illumination in images.

[6] False structure introduced by an inappropriate assumption of common lighting conditions across illumination boundaries

3.5 Experiment 4: Finding "significant" features in images

One of the claimed virtues of expressing data in terms of the PCA derived coordinate system is that, if each axis is scaled down by its variance, then because the axes are orthogonal and projections uncorrelated, the Euclidean distance from the mean becomes a reasonable measure of unlikeliness, the Mahalanobis distance (Duda & Hart, 1973). This is analogous to measuring a single value in terms of the number of standard deviations from the mean.

This could potentially label any sample by its scaled distance from the mean (in the PCA based coordinates) and hence a measure of the samples significance in the statistical sense. This possibly could be related to visual significance since based on the second order statistics, the samples furthest from the mean would also be the most information rich, (the most unlikely if from a multidimensional normal distribution, and hence informative). Given this, it was decided to explore this as a potential scheme for directing saccades to informative parts of the scene.

The full PCA (or LCPCA) expansion of a sample is both infeasible and undesirable. Given any independent noise, this will be concentrated in the lower components of low variance and by scaling by the variance, this will be amplified. Therefore only the first 15 LCPCs were used, these having been generated already. The following was then performed on an image to find areas of significance:

- The logarithm of the image was found and this was then high passed filtered: convolved with the high-pass-filter to reduce the effects of illumination, and making the image mean zero.
- Fifteen new images were produced, the convolutions of the transformed image with the first 15 LCPC's.
- For each of these 15 images, every pixel value was squared, then divided by the generating components variance.
- All 15 images were added, the pixel values divided by 15, and the square root taken.

This resulted in an image, the "brightness" corresponding to the scaled distance from the mean and two example results are shown in Figure 3.7. The system does label as significantly different most of the areas that subjectively would be of interest.

Figure 3.7: Two images and the corresponding "interest" of the various parts. The top right images was formed using LCPCs of size 32 by 32, the bottom right by components of size 64 by 64. The high "interest" at the bottom right corner of the top picture was caused because the convolution software assumed periodic boundary conditions.

Unfortunately this is all that can be said: that the system produces reasonable results. This system differs from standard models for the generation of saccades in using not just low frequency information but edge information as well, there is limited evidence that this is occuring in human vision (Findley personal communication). Without more work, little more can be said.

3.6 Experiment 5: Comparing orientation tuning curves

The orientation tuning curves of two "bar detector" LCPCs (component numbers 4 and 5) were calculated and compared both to those found found psychophysically in (Foster & Ward, 1991).

3.6.1 Results

As can be seen in Figure 3.8, although the match to the psychophysical data is not quite as good as found with variance normalisation PCA, given the limited set of images

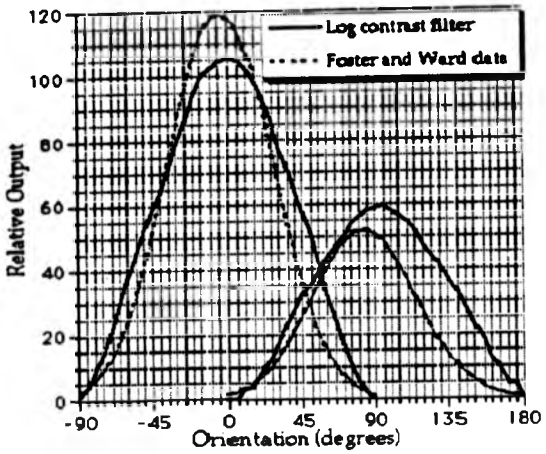

Figure 3.8: The orientation tuning curves of the two oriented "bar" detector (components number 4 and 5 solid line) as compared to the data of Foster and Ward (1991) (dashed line).

it is still impressive. The preprocessing is quite different, but the result approximately still holds.

3.7 Summary and general conclusions

If image interpretation rather than image representation is attempted, the features that are used should reflect the reflectance structure of the world, rather than just the incoming light which is dominated by the the effects of different illumination. PCA is a method of feature extraction that has a strong theoretical basis, unfortunately when applied to raw grey levels the results are dominated by lighting conditions and the effects of the structure of the world are masked. By using a related statistical method based on ratios: LCPCA, the feature extraction becomes lighting invariant in a way that doesn't amplify noise as variance normalisation of the inputs would.

The removal of the noise introduced by the previous variance normalisation technique results in a set of components that is more easily interpretable in terms of the analytically derived components of transformationally and rotational invariant correlations structures. Interpretation in this framework shows that with natural images, more of the image variance is accounted for by an orientation expansion of the image, rather than a spatial scale expansion. Although not directly comparable since the cortex uses

a non orthogonal expansion, this provides a possible insight into the physiologically and psychophysically observed additional resolution for angle rather than spatial frequency.

If PCA is intended as a metaphor for the generation of the representation in cortex, then these feature extractors will not be working on the raw intensities but on intensities transformed by retinal preprocessing. Exact LCPCA requires unrealistic preprocessing of the individual samples, but can be well approximated by performing simple PCA (with its simple Hebbian implementation (Foldiak, 1992)) not on the raw grey levels but on images transformed in a manner similar to that achieved in the retina.

Lastly the observed match of the orientation tuning curves of PCs to human psychophysical data was replicated for the LCPCs. This lends weight to the match not being caused by some artifact of the exact preprocessing used, but by the combination of "optimal" feature extraction and the anisotropic statistics of natural scenes.

The information theoretic framework which both PCA and LCPCA come from has drawbacks. The goal of vision is not to represent the visual input in a more and more compact form but to provide a representation that allows decisions to be made and actions to be performed. But with the extreme credit assignment problem inherent in low-level vision: does my success in getting food allow me to reinforce vertical bar detectors; information rather than decision theoretic criteria may be very useful. The expressing of the input in a form that makes redundancies explicit and allows efficient representation of inputs may well be useful in low-level vision where a more explicit training signal is not present. The success of the information theoretic framework in providing not only qualitative but quantitative matches to data: not only the orientation tuning curves as shown here, but to the colour sensitivities of ganglion cells (Atick *et al.*, 1990) and the contrast sensitivities of cells under different illuminations (Atick & Redlich, 1992), adds weight to this hypothesis.

3.8 General Conclusions on PCA and LCPCA

If little information is available as to the behavioural significance of visual input, how can the statistics of the input be used to learn? The previous two chapters presented one method of using the input to determine the representation: redundancy reduction and specifically PCA. In the previous two chapters, what evidence was presented that a process like this is used in the brain?

The first result of PCA: that the resulting operators resembled various ordered directional derivatives, is not unique to PCA. A number of other justifications for

approximate directional derivatives have been proposed that have nothing to do with the signal. Two such proposals are the minimum joint entropy argument for Gabors, and the differential geometry argument for JETS (Koenderink & van Doorn, 1987).

If when analysing some signal, both the frequency and the location of events are of interest, then because of problems similar to those of Heisenburg's uncertainty principle, an exact estimate of the location means a vague estimate of the frequency and vice versa. Gabor showed that for a one dimensional signal, the optimal representation as a compromise between the spatial and frequency resolution is a Gabor. Daugmann generalised this concept to three dimensions, and showed that the optimal filters, in terms of the greatest simultaneous resolution in the frequency and spatial domain, were sines and cosines multiplied by a Gaussian. The receptive fields found in cortex resemble some of the early PCs, but more closely they resemble Gabors. This could possibly be because the brain is also trying to optimise the joint spatial/frequency resolution.

JETS propose the role of receptive fields is function approximation (Koenderink & van Doorn, 1987). The image is seen as a two dimensional surface and the techniques of differential geometry are used to approximate this surface. Specifically, the receptive fields are used to perform a Taylor series approximation of the signal, and consist of directional partial differentials of different orders. These operators are smoothed with a Gaussian. For instance if the desire is to represent the signal with a second order expansion, then the six receptive fields would be G G_x G_y G_{xx} G_{xy} G_{yy}, where G is a Gaussian, and a subscript represents a partial differential in that direction (G_{xxy} represents the second differential of a Gaussian in the x direction, the first in the y direction.) The general form of the receptive fields produced by this scheme is very similar to those produced by PCA. For LCPCA all 15 can be approximately captured as different ordered JETS. The general form and the (very) approximate similarity to the receptive fields found in cortex, is therefore not a unique prediction of PCA.

As well as the general form of the components, the PCA's matched the Foster and Ward data. Components 4 and 5 were strongly aligned to the horizontal and vertical (as were the other components), and the receptive fields for the two bar detectors were different (and the difference in orientation tuning curves matched the psychophysical data.) Both these characteristics are much more difficult for a system where the representation is determined by constraints other than the signal it is to represent.

With JETS, two "bar detectors" exist (G_{xx} and G_{yy}) but the directions that the differentials are taken in is completely arbitrary. Traditionally the operators are aligned to the vertical and the horizontal axes, but this is purely convention. Nothing in the

theory specifies that it is vertical and horizontal that are important. In the Foster and Ward data, these axes have privileged status. The same holds for Gabors, although the vertical and horizontal axes are possible directions of analysis (and often used through convention), there is no reason motivated by the resolution for this to be this way.

The problems for Gabors, JETS, and other representations which are generated independently of the signal, becomes worse when the anisotropy is considered. In both, you would predict the analysis to be identical in both directions. Not only is the measured anisotropy between horizontal and vertical matched by PCA, but also the degree of anisotropy between horizontal and vertical. For any representation created independently of the world statistics, these differences must be difficult to explain.

Redundancy reduction in general, and PCA (or LCPCA) in particular, will not be unique in the prediction of a representation firmly aligned to the vertical and horizontal, and an anisotropy for these two directions. Because of the action of gravity, there will be more horizontal edges (because of horizons, man made objects), and vertical edges (again man made objects), therefore many representational schemes that use the visual environment to tune its representation, will be firmly aligned to the horizontal and vertical. Because of foreshortening, there will more horizontal edges than vertical edges (see chapter 4) and an anisotropy would not be unexpected. Therefore it is not unlikely that in a representation-scheme tuned to the visual diet, the representation will be aligned to the vertical and horizontal, and more resolution would be given to given to horizontal lines. PCA is one such scheme, and the exact match to the anisotropy is possibly unexpected.

The Foster and Ward (1991) is very well matched by the PCA of natural images, and the data presents problems for any representation-scheme generated independently of the the signal, but PCA is probably not unique, and other signal-based representation-schemes will also probably match this data. The one message is not that redundancy reduction, PCA, or any particular method is being used, but that whatever the method is, it seems likely that it is refined, learned, or defined in terms of the signal is has to represent.

Chapter 4

The correlational structure of natural images

Summary

An analysis of the spatial grey level correlations in a large number of different images was carried out. The structure of the correlations is understandable in terms of three different effects: self–similarity across the ensemble of images, foreshortening, and a predominance of horizontal and vertical structure. These correlations are first used to provide insight into previous work on the principal components (PC's) of natural images (Baddeley & Hancock, 1991). The correlations are then related to models of topographic map formation. These topographic mapping networks work by using spatial correlations in the input to create a new representation that captures the topographic relationship between inputs. Therefore any distortions in input statistics should be reflected in the final output representation. A comparison with previous psychophysical results provides a strong match to the theory proposed. Lastly an analysis of the uncertainty of the representations caused by both noise and limited sampling is investigated. The variability in the representation of distance is shown to be proportional to distance: this function is similar to Weber's law.

4.1 Introduction

In this chapter, an attempt is made to understand one of the simplest visual regularities, the way that measurements made at one location in the image are predictable from others, that is, how image grey levels are correlated. Although this is a simple regularity,

it can provide insight into the operation of various mechanisms, explanations of various psychological phenomena, and potential tests for various physiological models.

In unsupervised learning, linear models of feature extraction have been proposed such as principal component analysis (PCA) (Foldiak, 1992). Even in some non linear neural network models, their behavior can be understood in terms of a linear approximation (MacKay & Miller, 1990a; Miller, 1990). In such models the resulting features are determined by the input correlations.

For adequate spatial orientation an animal must also maintain a representation of retinotopic space. Physiologically this is analogous to the cortical "maps" of retinotopic space found in most higher animals. Although these are probably initially formed using genetically predetermined chemical gradients, fine tuning requires correlated input. Computational models have been proposed which perform this recalibration using locally correlated activity in the input, and if there are distortions in the input statistics caused by the nature of the visual world, then these will be reflected in the resulting representations generated by the mapping networks and hence the representation of retinotopic space. This can be tested for psychophysically.

Further, if a system operates, not with fixed structure, but constantly recalibrates itself based on statistical measures, then this system will display an uncertainty common to all statistical processes. This uncertainty will translate into an uncertainty of position estimates made by the system and hence into thresholds for distance discrimination. Again, this should be testable empirically.

4.2 Previous work

4.2.1 Statistical analysis of natural images

Gibson (1979) was the most famous exponent of the view that understanding the structure of the visual environment is necessary in order to understand perception. Field (1987) carried out a study of the spatial scale characteristics of six "natural" images. This spatial frequency analysis is related to the image correlations (as measured here) via the Wiener-Khinchine theorem (Connor, 1982). The analysis was of a limited set of six images (all of trees and rock), and the analysis averaged over all orientations, hiding much of the structure. Interestingly, the two power spectra in Figure 7 of the paper (see Field (1987)), shows pronounced vertical and horizontal structure. The power spectra revealed a $\frac{1}{Frequency^2}$ dependence of energy versus spatial frequency and this was used to justify as alternative to the standard Gabor front end a log Gabor pre-processing

stage. Later Atick (1992) used these characteristics to explain contrast sensitivity in the retina at various different levels of illumination.

The correlations present in one natural scene were also analysed by Srinivasan, Laughlin and Dubs (1982) but the scene was unrepresentative (a reed bed), and the measure of correlation did not take into account the mean level of luminance [1]. The measured correlation together with estimates of the signal-to-noise ratio of neurons was used to accurately predict the receptive field characteristics of fly retinal ganglion cells.

4.2.2 Models of the representation of retinotopic space

Psychological phenomena

Andrews (1964) raised the problem of finding the inverse cortical transfer function in order to create veridical representations of space. He proposed the use of various statistical regularities in the visual world to constantly recalibrate spatial representations, and claimed that a number of visual illusions could be explained in terms of the use of these regularities. Illusions of distance could be understood in terms of the mechanism used to set up and recalibrate representations of space. The proposal was that this constant recalibration of the representation of space is required to cope not only with damage to the system, but also with the changing geometry of the eye, distortions caused by the wearing of glasses, or the changing of head shape with age.

Craven and Watt (1989) showed that this approach lead to reasonable predictions involving a distance illusion. Ross (1990) showed that children brought up in different environments (such as the flat Fenlands, Scottish Islands, Inner city) and hence with different input statistics showed different length illusions, specifically different horizontal-vertical illusions.

Also relevant to this study is the observation that distance discrimination thresholds are usually proportional to the distance to be judged (the famous Weber's law (see Levi and Klein (1992) for a review). These discrimination thresholds possibly arise from an uncertainty in the position estimates.

[1] A large amount of confusion in the literature centres on two different definitions of correlation. In engineering, correlation often refers to a normalised measure of $\sum xy$. In statistics, the measure of correlation takes into account the mean level of the signal and refers to a normalised measure of $\sum (x - \bar{x})(y - \bar{y})$ and therefore is not effected by the absolute value of the signal. Srinivasan et al (1982) used an engineering type measure. In the present study, the statistical definition of correlation is used.

Physiological phenomena

Work on the formation of topographic maps is also relevant to the view that the metric of visual space reflects the way that it is set up and recalibrated. Topographic maps consist of representations in the brain, the structure of which reflects the structure of the input domain. Representations of visual space are organised spatially, and auditory representation is organised in terms of frequency. The way that these systems develop may help to explain the way that psychological representations develop.

Although the initial structure of these maps is probably specified genetically via chemical gradients, fine tuning of these maps depends on correlated input activity. There is evidence that there is constant recalibration of these maps, both from the spectacular regeneration in fish and amphibians after gross lesion of the inputs (Fawcett & O'Leary, 1985), and the less spectacular phenomenon of recalibration of representations after local input lesions in cats and primates (Gilbert & Torsten, 1992) Evidence of this recalibration can be found on the time scale of minutes and complete recalibration within months. The system responsible for this recalibration is dependent not on the inherent structure of the representations, but on the characteristics of the input. For example visual input redirected via MGN[2] to the auditory cortex results in the formation of a topographically organised representation of visual space, not a representation of pitch.

Computational models

Hebb originally proposed a theory of the "learning" of space perception (Hebb, 1949). He proposed that one principle would allow representations to self organise- that initially large numbers of possible connections exist, and that a given path is facilitated by use. He then proposed that eye movements are responsible for locating important features and acting as a reference frame. Unfortunately it is far from obvious if or how such a system could work and hence is of mainly of historical interest.

Willshaw, Prestige, and Von der Malsburg constructed a number of models inspired by different proposals for the induction of topographic maps in the optic tectum (Prestige & Willshaw, 1975; Willshaw & von der Malsburg, 1976; von der Malsburg & Willshaw, 1977; Willshaw & von der Malsburg, 1979). Moreover, the model were specified sufficiently that they could be implemented on computers, allowing their predictions to be bested. Kohonen also investigated topographic mapping networks (Kohonen,

[2]The MGN is the part of the lateral geniculate nucleus that is devoted to auditory input, much as the more studied LGN is the the thalamic relay for the visual cortex.

1982) but with more emphasis on engineering applications of which there have been a great number. Lastly Amari and co–workers have used tools from statistical physics to understand the behavior of such nets (Amari, 1980).

Although all are different in the details, they all operate on a similar principle. The basic idea is that there is some quantity in the pattern of the input that is more similar for nearby inputs than distant ones . Usually it is proposed that the activity of nearby units is more likely to have arisen from some common feature of the input, and therefore closer inputs should be more correlated (alternatively they are labeled with similar chemical markers). The networks operate by exploiting this correlated activity in the inputs by a two stage process. The dynamics of the output layer are constrained so that nearby outputs are correlated and more distant outputs are forced to be anti-correlated. These representing units are then connected to the input via modifiable weights that learn instances where the dynamics make the unit a "winner". This induces competition to represent the inputs, but forces nearby units to represent more "similar" inputs, and distant units to learn dissimilar inputs. Because of the spatial correlation in the image, nearby inputs are more similar so the network is stable when it is in topographic correspondence with the input (providing problems involving local minima are avoided). Network models working on these principles have also been used to model the physiological phenomenon of ocular dominance (Goodhill, 1992), and the development of orientation maps (Obermayer & Schulten, 1990).

4.2.3 The plan of this chapter

Firstly, two models will be outlined, one of feature extraction, and one of the formation of the representations of retinotopic space. It is shown that the results of these models depend critically on the correlations present in the visual world.

Next, the pixel intensity correlations are then estimated empirically from a large collection of natural images. It is shown that by averaging over large numbers, despite all the images being very different, a simple statistical structure emerges.

This statistical structure is then investigated. Firstly the decay of correlation with distance is modeled. This decay is shown to be lawful and determined by the fact that we view the world at many different distances. Its form explains the ordering of the "features" produced by proposed neural network models, and this form, combined with a model of topographic map formation, also explains the observed distance discrimination thresholds.

The structure of the correlations at different angles is then investigated and shown

to be understandable in terms of simple physical causes. This anisotropic structure is again related to feature extraction networks and provides an explanation of the close match of these models to psychophysical data. The structure of these statistics is shown not only to provide qualitative explanations of observed distance distortions, but also to provide accurate quantitative predictions of the size of these distortions in children raised in different environments.

4.3 Models

In order to show the potential relevance of the statistical analysis of the natural images, two models are outlined: one of automatic feature generation, and one an idealisation of neural networks for topographic mapping.

Principal component analysis

In chapters 2 and 3, we performed principle components analysis on natural images (PCA). The results of PCA are purely determined by the input correlations. We inferred earlier that the results we found were probably due to the anisotropic nature of the input statistics. By measuring these statistics directly, it should be possible to better understand the results found in that previous section.

Formation of the representation of space

As described previously, the brain creates and constantly recalibrates maps in topographic correspondence with the outside world, and physiological evidence indicates that correlated input activity is used to fine tune this representation. Again, a number of models have been proposed to account for the development of these maps, all varying in detail but all similar in that they rely on locally correlated input activity to produce topographically organised representations.

Rather than opt for any model in particular (most can only deal with localised input activity), an abstraction is proposed. Given that the only way these systems "know" that two measurements are close or far apart is that the activity of these measurements is highly correlated or uncorrelated, then we assume that any two measurements made from the world of equal correlation are treated by the system as being the same distance apart. Stated more formally:

Any pair of measurements made from the world that are of equal correlation

are represented as being an equal distance apart in the representation of these measurements.

This role of correlation as the currency of the brain is not new (Singer, 1990). To directly quote Goodhill on the relationship between correlation and distance:

> In order to identify the model derived $\cdots$ it is necessary (Yuille & Lee, 1991) to explicitly identify distances with correlation: in particular their analysis suggests that the correlation between x_i and x_j should be $\kappa - |x_i - x_j|^2$, where κ is a constant. This agrees with the interpretation of distances as correlations we have adopted here(Goodhill, 1992).

This is all that is required, given correlations that decay with distance, to create a topographically organised representation. Previous neural network theories can be seen as implementation level descriptions of this algorithm (in the sense of Marr (1982)). Rather than study the implementation level, we can study the system at the algorithmic level.

By empirically finding the relationship between correlation and distance,we can test the plausibility of these models and generate potential predictions. If the correlations decay both isotropically and equally across the visual field, then this will generate a veridical representation. If the relationship between distance and correlation is not isotropic, and the correlations decay at a different rate in different directions, then a system working on the assumption of equal correlations representing equal distances will induce distortions in the representation of distance. This should be testable experimentally.

Correlations and distance discrimination thresholds

If the metric used to calibrate representations of retinotopic space is correlation and the measured correlation has noise or some source of variability, then variability and hence uncertainty in the representation of retinotopic space will be caused. If a system with this uncertainty is used to estimate distances, then this will induce a limit to the accuracy of distance discriminations.

More explicitly, we have a function F relating correlation ρ to distance d: $d = F(\rho)$. If then the correlation has some uncertainty $\Delta\rho$, we want to be able to calculate the uncertainty in distance Δd. To do this we make a first order Taylor series expansion about d:

$$d + \Delta d = F(\rho + \Delta\rho) \approx F(\rho) + \Delta\rho F' + 0(F'') \tag{4.1}$$

Therefore to calculate the uncertainty in the estimate of distance Δd, all we need is an estimate of the function F relating correlation to distance, and an estimate of $\Delta\rho$, the uncertainty in the estimate of correlation. F can be found empirically from the data. The uncertainty in estimating correlation ($\Delta\rho$) could come from internal noise, or from the variability induced by a non stationary environment. The former can be modeled as a Gaussian noise source, the later can be empirically measured from the variability of correlation across different environments. The form of the errors induced by both will be found.

4.4 Computational Experiments

4.4.1 General Methods

The images

The images consisted of a collection of pictures of natural scenes. Pictures were taken on a 35mm camera with a 50mm lens (unless otherwise specified). The photographs were then digitised using a Hewlett Packard Scanjet Plus at the 75 dpi setting. The central 256 by 256 pixel region was then sampled and the grey levels normalised so that the lowest value was 1 and the largest was 256. No attempt was made to account for non linearities in the processing. This process was used to create the following 81 images:

- 15 images taken from the office environment consisting mainly of office scenes (people at work, furniture, corridors etc). This image set was known as the indoor subset.
- 15 images from the flat flood plain surrounding Stirling including river scenes, fields and country roads. This set was known as the country subset.
- 15 images taken within the city centre of Stirling including the insides of shops, super markets, and city housing. This set was known as the city subset.
- 36 miscellaneous scenes. These came from a large number of previous experiments and included pictures of text, faces, mountain scenes, cars and sculptures. Some of these images were taken with a 35-70 zoom lens making exact focal length unknown. This set was known as the miscellaneous subset.

Within the first three sets, it was attempted to remove effects of composition other than ensuring correct vertical alignment by the following techniques: The camera view finder was not used, the camera was pointed at random, and pictures were taken at set time intervals. This procedure produced three sets of 36 exposures. Out of focus pictures were removed together with pictures including lens obstructions. Fifteen from each set were then selected by shuffling and selecting the top 15. Equivalent angles were calculated by taking a picture of a ruler at a distance of 1 meter then digitising it with the other images. From this, it was calculated that $1° \approx 10.9$ pixels.

Correlation of images, or correlation of samples from images

There are two potential meanings of "the correlation within samples from a collection of images". This could be concerned with just as the statistics of the images. In this case each image could be considered as an "independent" sample from the world, and we can calculate useful things such as the variability of the correlation between images. Unfortunately this is not the quantity that we are interested in. What we are more interested in is not the statistics of images, but the statistics of these images as sampled by the eye, and this is confounded by the fact that a person would sample a single image many times.

To get the statistics of images as sampled by the eye, we would need to sample the eye movements as well as the images seen. Equipment in the advertising industry is available for doing such thing, but is incredibly expensive. An alternative is to study a creature with few eye movements (such as a cat), and use a head mounted camera. The samples seen by the eye are not independent: we look in the same area for a length of time and sample many times from different parts of a scene. Unfortunately we have only access to the scenes, not saccade locations. If samples cannot be treated as independent, he simplicity of the question of the statistics of natural images is lost, and we have to make somewhat arbitrary decisions about where in an image such samplings would occur.

The choice made here is that the action of saccades is to fixate all locations within every image. Unfortunately this somewhat arbitrary decision will both make a fairly large contribution to the statistics, and also render any estimates of such things as the variability of the estimate of correlation unreliable as these are dependent on the number of times we view a single scene.

This, combined with the fact that both the choice of images is somewhat arbitrary, means that quantitative results (such as the exact ratio of the correlation in the ver-

tical to horizontal) should be interpreted with caution. Also measures of variability can be misleading. The variability in correlation caused by sampling could be much less than the variability caused by choosing one particular image set as opposed to another. Despite this, to the extent that the assumption that all locations are fixated is approximately correct, we we can obtain qualitative and quantitative results of interest.

The calculation of the correlation structure

To calculate estimates of the spatial correlation within samples from a collection of images, the following technique was used. Every possible 64x64 pixel sample from the image set was taken (for each image this resulted in 36864 samples). For each sample the centre pixel value was found (33,33), and the relevant values measured to allow the calculation of the correlation of all other pixels with it. The correlation (Pearson's product-moment), calculated from all the samples was calculated using the following formula:

$$\rho_{x,y} = \frac{(N\sum_i (C_i I_{x,y,i}) - (\sum_i C_i)\sum_i (I_{x,y,i}))}{(\sqrt{(N\sum_i I_{x,y,i}^2 - (\sum_i I_{x,y,i})^2)) \times (N\sum_i C_i^2 - (\sum_i C_i^2))}}$$

where C_i was the pixel grey level value of the centre pixel within sample i, N was the number of samples, and $I_{x,y,i}$ was the value of the of the pixel at location x,y again in the sample i. This resulted in an "image" of size 64 by 64, with each location containing the value of the correlation of that point in the sample with the centre. This direct estimation technique avoided edge effects that cause problems in FFT (fast Fourier transform) based techniques.

4.4.2 Experiment 6: The general correlational structure

Method

The entire image set was analysed. Every possible 64x64 sample (that did not cross the edges) from all 91 images was found. This resulted in $(256-64) \times (256-64) \times 91 = 3,354,624$ samples. From the these samples a new 64x64 "image" was calculated with the value of a pixel at a given location being the correlation within the samples, of the measurements made at this location and one made at the centre pixel (location 33,33). Therefore the value at pixel 33,33 was always 1 (since the value at the centre was always perfectly correlated with itself, and the value at location 44,17 was the correlation within the samples of the pixel measurement made at the centre with the pixel 11 pixels ($\approx 1°$) to the right, and 16 pixels ($\approx 1.5°$) above.

For ease of visualisation, slices through the correlation were also found. Starting at

the centre, three lines were plotted through the correlation image. The first started at the centre (33,33) and continued to the top of the image (33, 64) and was known as the vertical slice. The second started at the centre and went to the right in a horizontal direction to location (64,33) and was known as the horizontal slice. Lastly a 45° slice was made in the north west direction to location 64,64 and was known as the 45° slice[3]. Along these slices, the correlation was found resulting in estimates of the correlation versus eccentricity in these three directions.

Results

Figure 4.1 shows the results of the correlations calculated for all the images. Note that the correlations are remarkably regular and decay away monotonically from the centre. in fact each of the samples has quite complicated but different structure. By averaging over the samples, what is left is the simple and intuitive observation that measurements made close together are on average more similar than measurements made at a distance.

Apart from the monotonic decay of correlation with distance, there are two other notable aspects of the the correlation. Firstly correlation decays faster in the vertical direction than the horizontal: the correlations appear squashed. Secondly there is also limited evidence for correlations extending further in the vertical and horizontal directions than would be expected just from the effect of the squashing (in the slices through the correlations [see figure 4.1], despite the squashing of the correlations, they extend at the same rate for both vertical and 45° slices).

Discussion

As can be seen, although the individual images are very different, the picture that emerges from the average correlation has a much simpler structure. The correlation of the ensemble has three kinds of obvious structure: 1) they decay monotonically with distance, 2) they extend further in the horizontal direction than the vertical, and 3) they are more pronounced along the cardinal axis. The form and cause of this structure is investigated.

4.4.3 Experiment 7: The decay of correlation with distance

One major structural feature of the correlations is that they decay with distance. The form of the decay is fitted with a number of functions and the cause of this decay

[3]Because of the square sampling geometry, this slice was $\sqrt{2}$ longer.

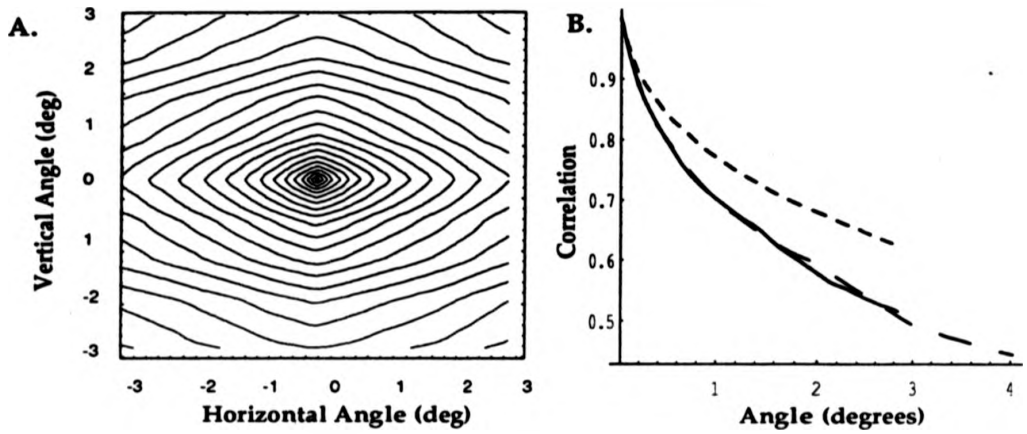

Figure 4.1: The correlations present in the entire image set. **A** shows a contour plot showing the structure of the correlations **B** shows three slices although the correlations. The solid line is the correlations in the vertical direction, the short dashed line the correlations in the horizontal direction, and the long dashed line the correlations at 45°.

hypothesized.

Method

The decay of correlation was measured over larger angles than previously, The variability of the correlation over different image subsets was also measured (so that the variability could be used in the function fitting). This was done by repeatedly finding the correlation in subsets of the image set. This allowed the calculation of both the average correlation and the relative variability in correlation at different eccentricities.

More precisely, from the entire set of images, a subset of 8 random images was found. From these 8 images, 400 random horizontal samples of length 129 pixels ($\approx 12°$) were taken. These were used to calculate the correlation of each pixel in the sample with the one furthest to the left within the sample. This process of subsampling 8 images was repeated 40,000 times, and from the resulting 40,000 correlation estimates, the average correlation and the standard deviation of the correlation estimates was calculated. This process was repeated with the samples being taken in the vertical direction.

To characterise the form of the decay, four power type functions were fitted of the form $1 - \rho = f(\theta)$. The criterion used was the squared deviation at each point on the graph, with this divided by the empirically found standard deviation of the correlation

at this θ. The first point at angle zero was ignored[4]. The power law type functions were:

- $1 - \rho = m \log(\theta) + c$
- $1 - \rho = m \exp(\theta) + c$
- $1 - \rho = c\theta^m$
- $1 - \rho = c\theta + m\theta^2$

where ρ is the correlation, θ is the angle (calculated using the 10.9 pixels equal one degree), and the fit was for the two parameters c and m.

Results

Figure 4.2 shows a horizontal slice through the measured correlations together with the four fits. Note that the plot is of distance against $1 - correlation$. Table 4.1 shows the parameters of the fit and the χ^2 of the various fits.

Note except for the early points, the decay with correlation is well modeled as a logarithmic function of distance. This is also true of other image sets experimented with, and if another free parameter is used ($1 - \rho = m \log(\theta + c2) + c1$), the fit is near perfect (see Figure 5.4).

Discussion

The correlations decay in an approximate logarithmic fashion. Why should this be? One possibility is that the individual images are fractal. Field (1987) claimed, based on six images, that images can be characterised as fractal, and that they they have power spectrum of the form power $\propto 1/\text{frequency}^2$. Whilst this may be true of the six images Field studied, this is not true of all images. The sky, faces, a wall, books, a book; all these images have a characteristic scale, and therefore these are not fractal.

One reason for this form of decay of correlation with distance is found could be for the following reason: Consider images consisting of random sized one dimensional

[4]The correlation at distance 0 is always one since it is the correlation of a measurement with itself. This is somewhat artificial because both the signal and the noise are forced to be correlated with each other and this is not the case for all other measurements. Ideally we would make two measurements with the camera of this central pixel value, then it would be possible for the correlation at 0 to differ from one. Since the fits are weighted by the empirical variance of the measurement, and this would be zero for distance 0, all fits would be constrained to pass through $\rho = 1, d = 0$. This is artificial and to avoid this problem of perfectly correlated noise, we removed this point from the calculation.

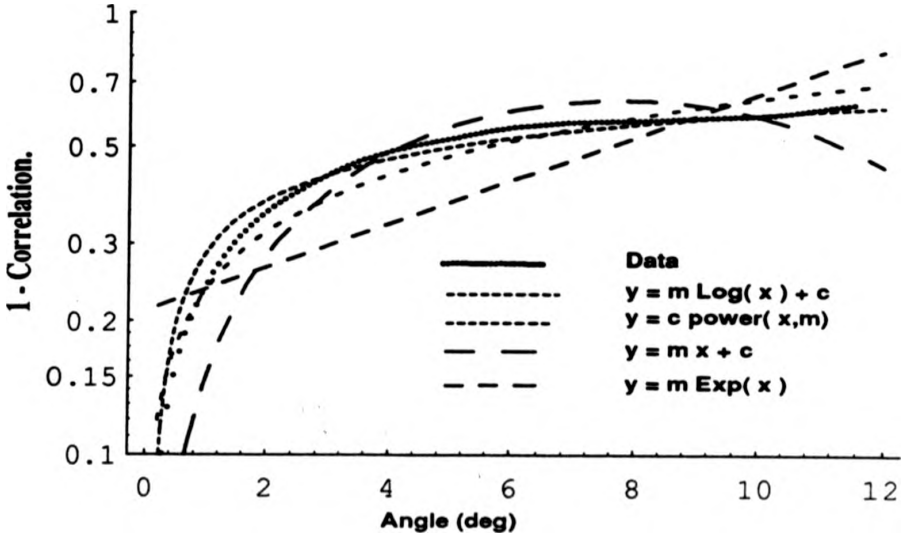

Figure 4.2: The correlations in the horizontal direction together with four power family fits.

Function	a	b	$Fit(\chi^2)$
$1.0 - \rho = a \log \theta + b$	0.128	0.287	10.82
$1.0 - \rho = a\theta^b$	0.237	0.426	24.10
$1.0 - \rho = a\theta + b\theta^2$	0.162	-0.0104	46.93
$1.0 - \rho = a \exp b\theta$	0.211	0.112	193.1

Table 4.1: The χ^2 fits of the four investigated fitting functions to the data in order of descending fit. Since the χ^2 values are based on arbitrary scaled standard deviations, the absolute value of the χ^2 fit is uninformative, only the relative value is of interest.

uniform patches. All measurements made with the patch are identical (and hence correlated). All measurements made beyond the patch are entirely random and hence uncorrelated. For an individual view of an patch, the correlation will be constant (equal to 1) for the size of the object, and zero beyond the object boundary. If now the patch is viewed from a different distance, or a different sized patched is viewed, then the correlation across both images will be the average of the correlation as found in these two images. To find what the correlation would look like in this case over a large number of images, the following computational experiment was performed:

- The size **X** of a patch was generated from a uniform probability distribution of width 10,000.
- Ten samples were generated from this distribution (from a uniform distribution range $\pm$ **X**).
- This process was repeated 10,000 times.
- The resulting probability distribution was estimated by normalising the distribution to standard deviation 1, and binning into bins of size 0.1 sd.

This process resulted in the probability distribution that would occur from adding a large number of different sized uniform distributions where the size of the distributions is also from a uniform distribution.

This distribution is shown in Figure 4.3. As can be seen, the resulting probability distribution decays as an approximately logarithmic function of distance. The discrete sized patch assumption is obviously a simplification. The experiment was repeated, this time the samples being drawn from a Gaussian distribution, (the standard deviation chosen again from a uniform distribution). Again the resulting distribution decayed approximately logarithmically.

To summarise, the result of adding large numbers of uniform distributions of width drawn from a uniform distribution, is a distribution which decays approximately logarithmically with distance. Translating back to our idealisation, if the world consisted of patches of uniform intensity[5], but with the size of each patch being random, and the intensity of neighboring patches being random, then the correlation will again decay approximately logarithmically.

This provides a potential explanation for the logarithmic decay in correlation. It is not that in any one single image or sample the correlation decays logarithmically, but rather that the images are viewed at many distances, and there are many different sized objects in them. Integrating over them generates the logarithmic structure that is found. The real situation is obviously much more complicated, but the essence is captured by this simple model of many different objects of many different sizes. This is done without the need to propose that images have self similar structure, or are fractal, and only requires that they are of many different sizes. This seems a much simpler explanation.

[5]That is all measurements within a patch being perfectly correlated

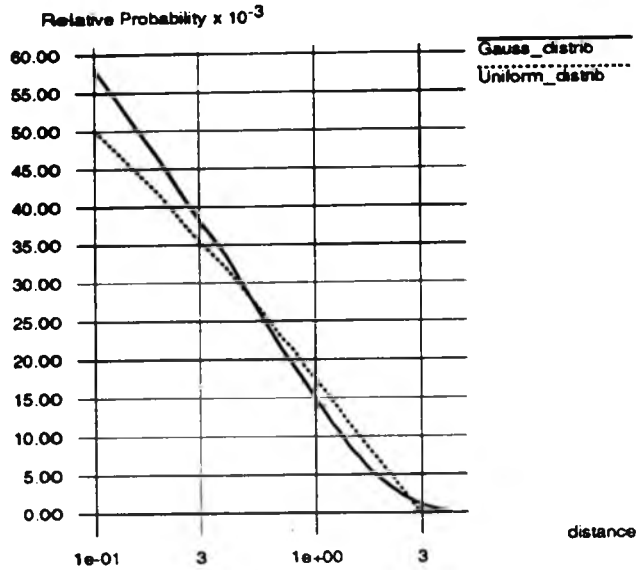

Figure 4.3: The results of combining a large number of numbers of zero mean distributions, each with a different characteristic scale drawn from a uniform distribution. The solid line is of summed Gaussian distributions, the dashed line is formed out of uniform distributions (see text). Both are straight lines when plotted on log-linear coordinates indicating that the decay is a logarithmic function of distance as found with the images.

4.4.4 Experiment 8: The shape of the correlations

As well as the decay of correlation with distance, the correlations possessed additional structure. The correlations extended further in the horizontal direction than in the vertical, and were "peaked" (extending further in the horizontal and vertical directions than in others). Experiment 8 attempts to identify both the nature of this structure and its physical causes.

Method

The correlations for both the country subset and the city subset were individually analysed using the technique described in the general method (section 4.4.1).

Results

The correlations for the country set are shown in Figure 4.4. The correlations for the city subset are shown in Figure 4.5. As can be seen, the correlations from the country set decay in what appears to be a squashed circular fashion (they are oval in shape). At least plausibly they could be due to the correlations originally being isotropic, then being subjected to an affine transformation.

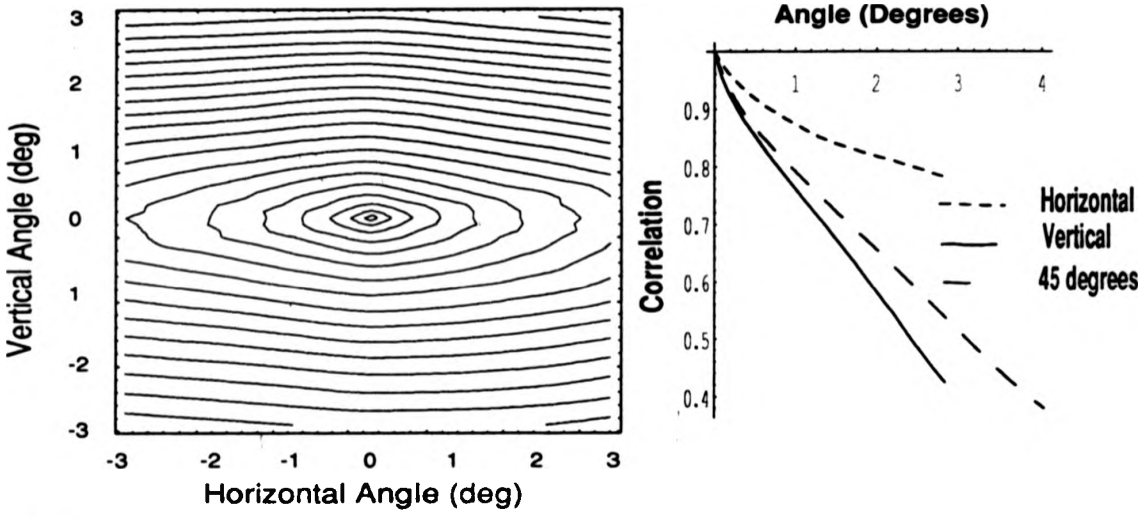

Figure 4.4: The correlations in the country image subset. Again the "image" is a contour plot of the correlations with the central point, and the graph represents the correlation as found in three slices through these correlations (in the vertical, horizontal and vertical direction- see experiment 6).

The correlations from the city subset are definitely not oval and could not have resulted from a simple affine transformation of originally isotropic correlations. The correlations do extend further in the horizontal direction as opposed to the vertical, but much more dramatic is that they extend much further in the vertical and horizontal directions rather than in other directions. The structure of these two images sets appears different and unlikely to be caused by a single process.

Discussion

Besides the decay with distance, a second component of the structure is the flattened nature of the correlations. The reason for this becomes clearer in the correlations of just

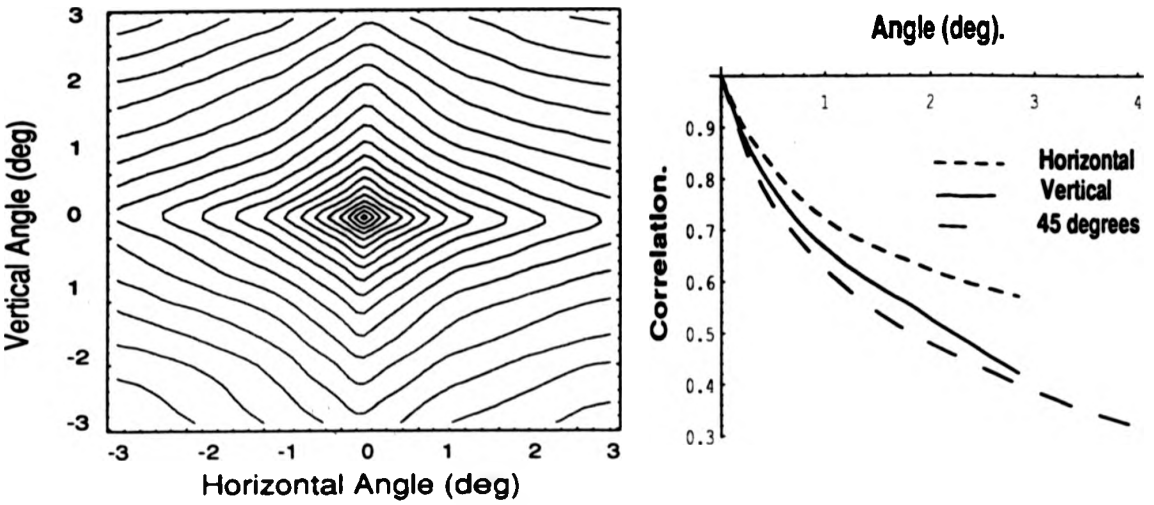

Figure 4.5: The correlations in the city image subset.

the country image set (Figure 4.4). This shows a much more pronounced flattening of the correlations. Within the country set, there are large numbers of landscapes with the texture of the ground plane being foreshortened. Foreshortening will cause squashing of the correlational structure in the direction of foreshortening, a fact used by some shape from texture algorithms ((Brown & Shvaytser, 1990) contains a more technical description of this process).

The reason foreshortening will cause the "squashing" of correlations can be understood in terms of a simple generative model. Consider again a world with patches of uniform brightness, this time with all the patches being of the same size as for example are the squares of a chess board. If the chess board is viewed from directly above, then the correlation will decay equally in the vertical and horizontal directions. If we now tilt the chess board away from us, then even although the squares (patches) are the same size, due to foreshortening they will appear wider than they are tall. The correlation will therefore decay quicker in the vertical direction rather horizontal, with the degree of tilt determining the anisotropy in correlation.

Most of the scenes in the country subset are foreshortened in the vertical direction. The correlations will therefore be squashed in this direction and by an amount dependent on the amount of foreshortening contained in the image set- wide open landscapes will produce highly anisotropic correlations. while those from within a house

will contain very few foreshortened parts (and these will be foreshortened in random directions), and will therefore produce much more isotropic correlation.

The second aspect of the structure of the general correlations was the peaked nature of the correlations in the vertical and horizontal directions. This is much more pronounced in the correlations that come from the city image set (Figure 4.5). Correlation in images is caused because measurements from an object are likely to be statistically related. This will be lessened by transitions between objects. Especially in man-made environments, there is a predominance of objects with vertical (lamp posts, tall buildings, doors), and horizontal structure (shelves, cars, pillars, horizons). This is reflected in the correlations from the city scenes and man-made environments in general (and some others such as forests), with objects aligned in the vertical and horizontal directions having correlations that are peaked in the horizontal and vertical directions.

These two physical causes provide an explanation both of the strongly aligned nature of the PC's and the relative orientation tuning curves of the two bar detecting PC's (in chapter 2). The PC's of isotropic fractals are randomly oriented, whereas the PC's for natural images are strongly aligned to the horizontal and vertical. Foreshortening provides an explanation of this, since it causes far more variability in the vertical than the horizontal direction. In a system that is trying to account for as much variability as possible, accounting for the vertical variations is most important. In terms of PCA, the second principle component is thus vertically oriented, rather than horizontal.

Foreshortening is also the cause of the observed difference in orientation tuning curves of the bar detectors found in Chapter 2. In that chapter, it was shown that the difference was caused by the anisotropy in the correlations. If this was caused by foreshortening, one prediction is that if it is the environmental statistics that are causing the difference in orientation tuning curves, then children brought up in rural and hence more foreshortened environments should show a larger difference in orientation tuning curves than those raised in the (less foreshortened) urban environment. This has not yet been tested.

The observed distortion of the correlations also has relevance to the model of the formation of topographic maps. If equal correlation is represented as equal distance, and the correlations decay faster in the vertical direction than in the horizontal, then a set physical distance will be represented as larger in the vertical direction than the horizontal. This effect has been widely observed in humans subjects and is known as the horizontal-vertical illusion[6]. To generate quantitative predictions of the size of

[6]To observe the horizontal-vertical illusion, view a pen in the vertical and horizontal directions.

the horizontal vertical illusion, estimates of people's visual diets are needed. No child observes only city scenes or country scenes and in experiment 9 below, an attempt to use more realistic visual diets was made in order to compare the distortions in the correlations with observed distortions in human retinotopic space.

4.4.5 Experiment 9: Environmentally determined horizontal vertical distance estimation differences

Since the structure of the spatial correlations is determined by the environment, and it is hypothesized that distortions in these correlations would be reflected as distortions in the representation of space, the extent of distortion for different environments was measured in order to be able to relate it to observed spatial distortions in people.

The proposed theory states that any two points of equal correlation will perceived as appearing of equal distance. In order to make comparisons with the psychophysical data, the visual angle required to get a level of correlation of 0.9 was found in both the vertical and horizontal directions. The ratio of these two distances was calculated. The same process was repeated for correlation levels of 0.8 and 0.7 and the geometric mean was found. This mean ratio should, if the theory is correct, correspond to the ratio of perceived distances in the two directions.

Method

Two separate image sets were analysed. The first image set, intended to be representative of a Fenland child's environment consisted of the indoor image set combined with the country image set. The country image set consisted of a number of flat countryside landscapes typical of the Fenlands; this was combined with the indoor set to approximately mirror indoor life. The second image set again consisted of the indoor set, this time combined with the city set. This was meant to mirror Glasgow children's visual diet. Again the correlations of these two image subsets were found, slices taken in the vertical and horizontal direction, and the visual angle needed to get a correlation of 0.9, 0.8, and 0.7 in both horizontal and vertical direction was found.

Results

Table 4.2 shows the results obtained. The geometric mean of the estimates was used, since these are ratios. For comparison, results from a study on environmental influences

In most subjects, the pen appears longer in the vertical direction.

Correlation Level	Horizontal vertical ratio Simulated Glasgow environment.	Horizontal vertical ratio Simulated Fenland environment
0.9	1.30	1.48
0.8	1.31	1.56
0.7	1.33	1.66
Geometric mean	1.31	1.56
Experimental results	Real Glasgow environment	Real Fenland environment
Ross (1990)	1.27	1.58

Table 4.2: The predicted and measured horizontal vertical illusion for children in urban and rural environments.

on the extent of horizontal–vertical illusions are also presented. The match between the estimate from theory (1.3 and 1.56) is very close to that from the experiment (1.27 and 1.58).

Working out the "statistical significance" of this result is both difficult and probably irrelevant. The study by Ross quotes only standard deviations of the perceived vertical and horizontal lengths, not the variability of the ratios. The values found here are definitely within a rough "worst case" calculation of the variability of her results.

The ratios derived from image correlation have two sources of variability: sampling variability, and variability caused by the choice, and more importantly, the ratio of indoor scenes to the country and city sets. Sampling variability could be calculated, but unfortunately it is likely that the variability caused by the choice and ratio of images from the two different image sets will dominate. Without any means to calculate what a reasonable ratio is, or a means to sample representative images from these two environments, calculating the sampling variability seems irrelevant.

Discussion

Estimating the exact structure of the visual input is a difficult operation. The chosen mix of 50% indoor life to 50% outdoor is arbitrary, but without continuous sampling of a child's visual environment better estimates are hard to come by. It was also the first attempt. Despite this, the quantitative match to the psychophysical results is close.

Even if the image set is very unrepresentative, we must have a qualitative match. There must be a horizontal vertical distortion in the statistics because in any reasonable environment there will be more foreshortening in the vertical direction. Therefore

a system calibrating using correlations will show a horizontal-vertical illusion. The degree of foreshortening will also be much greater in the more landscape rich country environments than in the city ones. This would account for the larger illusion in the children reared in the country environment.

It is intriguing to note that the match is not only qualitative but quantitatively correct to less than 2% error. However, without estimates of the visual diet of the children that we can trust, and estimates of their variability, it is difficult to say anything more quantitative about the match (which is tantalising since the match is so close).

4.4.6 Experiment 10: The horizontal-vertical illusion at different angles

The combination of foreshortening, together with an increased predominance of horizontal and vertical lines, should result in the correlations decaying faster when the orientation of the measurement deviates slightly off vertical rather than exactly vertical. The extent of this effect is found and compared to previous existing psychophysical data.

Method

The correlations from the image set described in experiment 12 were found. Those correspond to a distance of 1.5° ($\approx$ 16 pixels) for all angles were found and the difference between this and perfect correlation (1.0) calculated. In terms of the theory where correlation is used to calibrate estimates of distance, angular variations will translate into angular variations in estimated distance. It is therefore interesting to see exactly how correlation varies with angle.

Results

The results of this operation are shown in Figure 4.6. Again note that a point in the horizontal direction has the highest correlation ($\pm$90°), the correlations is lower in the vertical rather than the horizontal direction, and the least correlations are at about 45° to 20° from vertical.

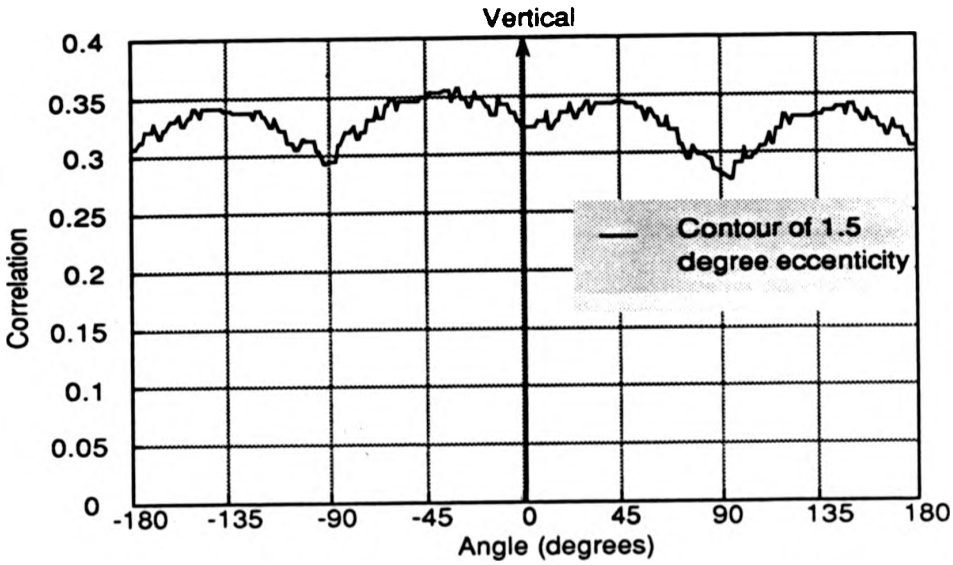

Figure 4.6: The contour of 1-correlation at a distance of 1.5° as measured for a number of different angles. Things to note from this diagram are that the correlations extend the least distance in the horizontal direction (±90°), further in the vertical direction, but furthest at an angle of $\approx \pm 30°$. Note also that the correlations are slightly anisotropic extending further to the left of vertical than to the right, and that the correlations at 45° are closer to those at the vertical than those in the horizontal. If correlations are used to calibrate the retinotopic representation of space, all these will cause measurable distortions in space.

Discussion

Inspection of the correlations reveals that if they were used to calibrate a representation, then other illusions than the horizontal-vertical would take place. The angle with largest correlation would be the horizontal: a line at this angle would appear shortest, but the angle with greatest correlation is not vertical, but about $20 - 45°$ away. As well as finding the existence of the horizontal vertical illusion, a number of workers have investigated the effects of orientation on the perceived size of the illusion (Cormack & Cormack, 1974; Underwood, 1966; Pollock & Chapais, 1952).

The first effect found is that the horizontal vertical illusion is not at its greatest extent at 90°. Cormack and Cormack (1974) tested a large number of stimulus configurations and found "*that in no case did the true vertical standard give the largest*

illusion.[7]". The largest illusion was usually found when a horizontal line was compared to one about 20° from the vertical. This was also found by Pollock and Chapanis (1952). This matches the data from the correlations.

Explaining a larger horizontal vertical illusion at horizontal vs $20° - 30°$ would cause considerable problems for models that rely purely on compensating for foreshortening. For such models, surely the greatest illusion would be found for a vertical line compared to a horizontal one, not a line $\approx 20°$ off vertical as found by Cormack and Cormack (1974). Again, the exact point of the maximum illusion cannot be reliably calculated because of uncertainty in the choice of image set, but that the maximum illusion will not be for a comparison of horizontal with vertical is reliable. Any image set that contains both foreshortened images, and images with more horizontal and vertical structure, will have the fastest decaying correlations either side of vertical. Despite uncertainty in the actual diet, it is reasonably certain that the visual diet of humans will have these characteristics.

4.4.7 Experiment 11: The relationship between distance discrimination threshold and distance

If correlation is used to calibrate the representation of retinotopic space, but the estimates of correlation are noisy, then this will cause distortion and variations in the representation of space. These variations will add random fluctuations to distance measurements, thereby limiting the distance discrimination resolution.

The effects of two kinds of noise in correlation on the thresholds for distance discrimination will be examined: simple Gaussian random noise, and noise from limited sampling. These noise models were used to estimate the relationship of distance estimation thresholds to visual angle (see equation 4.1).

Method

The relationship between visual angle and (a) correlation, (b) the variability in correlation, and (c) the differential of the correlation function was estimated for distances up to 6° in the horizontal and vertical direction using the following techniques:

- Eight images were randomly drawn from the complete image set. From these images, 400 one dimensional samples of length 7° were taken in the horizontal and vertical direction. These samples were used to calculate the correlation of

[7] Cormack and Cormack 1974 p 210.

image grey levels at different distances in the vertical and horizontal and vertical directions for this image subset.

- The above process was repeated 40,000 times. From the samples, average correlation as a function of distance was calculated. Also calculated was the standard deviation of the correlation as a function of distance.

- In order to calculate the uncertainty in distance estimates, we need an estimate of one over the differential of this function and this is unstable. Therefore we smoothed the estimated correlations using a second order moving average (each correlation estimate was replaced by the average of it with its neighbour).

- The differential of the smoothed correlation function was approximated as the difference of the correlation at one point and its neighbor. The inverse of this function was found, resulting in a number proportional to $\frac{\delta distance}{\delta \rho}$.

- To find the form of errors if the variability of the estimate of distance was based on limited sampling, the estimate of the variability of correlation was multiplied by the estimate of $\frac{\delta distance}{\delta \rho}$.

This resulted in the following estimated functions: 1) the function of correlation versus distance; 2) a function proportional to the standard deviation of the correlation estimates with distance; 3) one over the differential of the function of correlation with distance; 4) and the result of $\frac{\delta distance}{\delta \rho} \times$ the standard deviation of the estimates of correlation.

Results

The results of this computation are shown in Figure 4.7. The correlations decay approximately in the logarithmic fashion as found previously in both the horizontal and vertical directions (A) The inverse of the differential is approximately a straight line (B) If the variability in correlation is constant (caused for instance by constant noise), then the threshold for estimating a distance will be determined by this function and it will be an approximate linear function of distance.

The standard deviation of the correlation estimates is shown in C It is zero at zero distance (the correlation is always 1.0) and increases with distance. Somewhat surprising is that it is very similar in both the horizontal and vertical directions, even although the correlation at each distance is different. Lastly, the result of multiplying

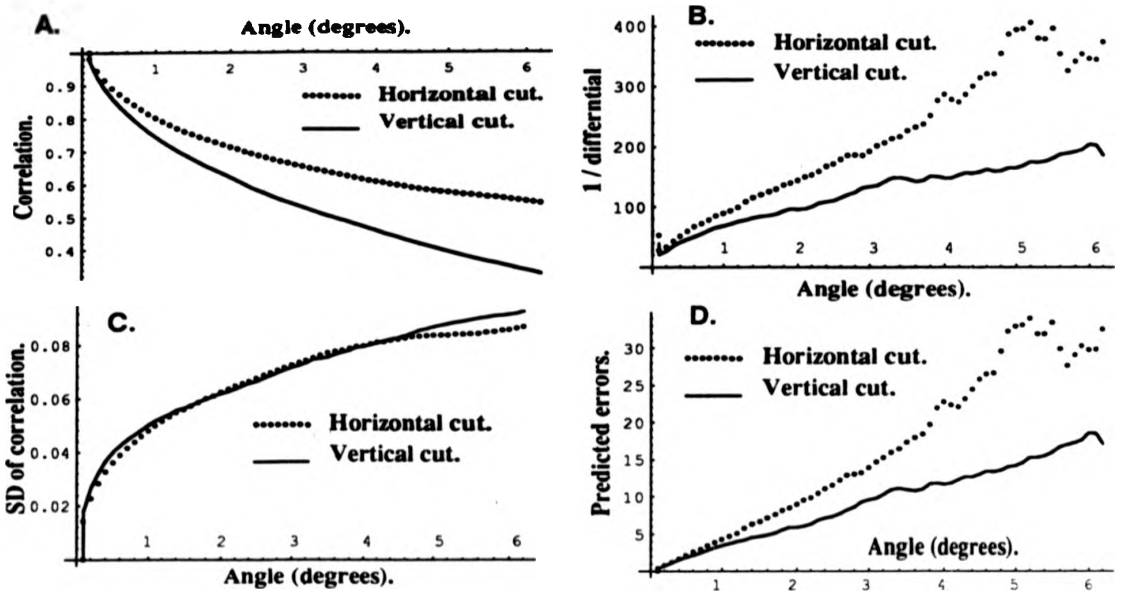

Figure 4.7: The estimate of the distance discrimination threshold. **A** shows the measured correlation in the vertical and horizontal direction. **B** This shows the estimate of $\frac{\delta distance}{\delta \rho}$: this would be the estimated threshold if the variability in correlation was caused by a set amount of noise. **C** This shows the estimate of the variability in the correlation found by repeatedly taking different samples. These were used in **D** to find the estimates of the threshold that would result from constant recalibration of the system where the estimate in the correlation will vary depending on the images seen.

the inverse of the differential of correlation by the variability of correlation is shown in **D** This will be the distance discrimination threshold if the noise in correlation is caused by estimating the correlation over a limited number of samples. Again the result is approximately a straight line.

Discussion

The most striking result is that the threshold for distance estimation for both sources of correlation variation is a linear function of the distance for a large range of angles. The reason for this is that $\rho \approx m \log distance + c$ so $\frac{\delta distance}{\delta \rho} \approx distance$. Even if the variability of correlation is not constant, this term dominates the predicted thresholds: the distance discrimination thresholds are determined by the form of the correlations much more than their variability.

This relationship of threshold being proportional to distance has been widely measured in human psychophysical performance and is consistent with Webers law (Levi & Klein, 1992). While a Weber–type relationship might be expected for such measurements as luminance estimation (possibly caused by receptor log compression), for distance estimation, it is harder to explain. If distance estimations are made by combining a number of intermediate measurements (with the errors on these measurements independent), then this would predict a square root law relationship. This has not been found. It should be noted that the error predictions are not just determined by the model but also by the statistics of the world. For animals raised in environments with different statistics, the predicted errors would be different.

4.5 Summary

4.5.1 Correlations in the world

The first portion of this study consisted of an investigation of the correlational structure of the world. The images analysed all had very different grey level structures. Despite these large differences, when viewed in terms of the correlations over the ensemble, a much simpler structure emerges. Identifying the physical causes for the three main structural features allows both an understanding and a limited predictive power for the structure of correlations in new and unsampled image sets.

The majority of the correlational structure can be accounted for by three different effects: 1) The correlations decay in an approximately log law fashion. This is not caused by the fractal nature of the individual images (some are fractal, others are not), but occurs because we are averaging over images viewed at many different distances and therefore the ensemble is by necessity self similar. 2) The correlations decay more slowly in the horizontal direction than the vertical: this is caused by the fact that the foreshortening in the images compresses the image statistics[8]. 3) The correlations decay slower than expected in the vertical and horizontal directions. This was suggested to be caused by the presence of a bias in the (man made) world for objects in these two directions.

[8]This has relevance to theories such Atick's that one role of retinal processing is to whiten the signal (make all spatial frequencies equal, and hence remove redundancy. Since the correlations are different in the vertical and horizontal direction, so will be the spatial frequency, and hence to whiten the signal the ganglion cells should have predictable anisotropic receptive fields. This should be testable.

4.5.2 Correlations and optimal feature extraction

The measured correlations also allow an understanding of the PC's derived from natural images. Because it is reasonable to assume that samples from images are stationary, the PC's can be understood purely in terms of the autocorrelation function. The decay in PC operator standard deviation obeys a power law, which is caused by the approximate power law correlations in the image. The operators are firmly aligned to the vertical and horizontal because foreshortening causes most variation in the vertical direction, and least in the horizontal direction. Lastly, the degree of anisotropy in the correlational statistics, and hence the anisotropy in the orientation tuning curves, is caused by the amount of foreshortening in the visual diet. The match to the psychophysical data relies on the image set used being representative of the human visual diet. If this account of the formation of orientation tuning is correct, then the measured orientation tuning anisotropy in cultures living in highly foreshortened environments should be larger. This has been found for distance estimation, but orientation has not yet been tested.

4.5.3 Correlations and the representation of space

Topographic mapping networks have been widely proposed for modeling many of the phenomena associated with topographic maps, and have provided a reasonable model for phenomena such as the development of ocular dominance stripes, the development of topographic representations, the development of the smooth variation of orientation sensitivity and the results of various deprivation experiments (see next chapter). Despite this, no serious psychological tests have been made on them. By abstracting away from the implementation (needed if they are to be tested physiologically), and concentrating on the algorithm they are implementing - that correlation equates with distance, it has been shown that they can generate psychological predictions.

The foreshortening of correlations, given their use to estimate distance, will cause a horizontal-vertical illusion. This will be robust across different environments and provides an implementable and parsimonious explanation of the psychological phenomenon. That it not only qualitatively, but also quantitatively predicts the size of the illusion in children raised in different environments, is striking.

Lastly, an analysis of the variability in distance estimates caused by variation in the correlation, showed that this would produce a system with uncertainty in distance estimates. This in turn leads to distance discrimination thresholds, which are linear

functions of distance. This is consistent with a Weber type law for distance. Any particular implementation may cause distortions in the representation of a particular form but one major and consistent source of error must be error in the calibrating signal. Thus what was previously a difficult problem (why do we show Weber type behavior in distance discrimination?) becomes evidence in favour of a system which is calibrating itself based on correlation.

Chapter 5

Correlation and the Functional Geometry of Striate Cortex

Summary

It is proposed that the mapping between the world and cortical coordinates in V1 matches that which would be expected if the cortex was calibrated using correlation. Specifically, a neuron representing a measurement made at a give eccentricity is represented in cortex a distance from the fovea proportional to one minus the correlation of the luminance measured at this point, and that measured at the fovea. Given an estimate of the the input correlations, this proposal predicts a specific mapping between the world and cortex. The correlation of image grey levels was estimated from random samples within a set of 72 images. This allowed the inferring of the shape of a map if this map was based on this correlation based metric. This inferred map was then compared to that measured in V1 in Macaque (Van Essen & Maunsell, 1984; Tootell *et al.*, 1982). The match of the two geometries was very good: 1) The form and the parameters of the cortical magnification function matched. 2) The ratio of cortical distance devoted to the representation of the horizontal to vertical meridians were very similar. 3) The form of the variability in the geometry of V1 within species is consistent with a correlation based metric. It was attempted to interpret this match in terms of the system being primarily interested in the point of fixation, and the average correlation with the point of fixation being a reasonable estimate of how relevant an eccentric measurement is to the identification of the foveated object. This interpretation avoids many of the problems of the approach of Schwartz who proposed that the form of

the representation is determined by required invariance properties.

5.1 Introduction

Vision can be seen as an active dynamic process. Rather than create a representation of the whole visual field and hope that areas of interest are well represented, a more flexible behavior is to identify the areas of interest and actively foveate them (Gibson, 1979). Although it is possible to direct attention to areas not in the fovea, for most of the time eyes are directed to the object of interest. If the identity of the object is still not clear, another saccade is performed revealing more information about the object's identity.

The standard representation proposed in vision (usually implicitly) is that of the image or a transformed version of it. In a passive computer based recognition system, the whole visual field can be scanned (theoretically in parallel), and features extracted from the whole image. The mathematical tools (convolution, histogram equalisation etc) are image based, and for a system that derives its information from a single view, a representation with a geometry that reflects that of the world is useful. From it distances can be easily found, and angles in different parts of the image easily compared.

Although this static form of representation is a useful tool to investigate perceptual mechanisms; observing your eye movements whilst reading this page, or walking, or looking around the room, is enough to convince most people that this is not the predominant mode of the visual system. For the majority of the time, if something is of interest, you do not observe it out the corner of your eye, but you fixate it. The low-level representations used in vision should be optimised not for the single all encompassing view, but for this active sampling of the image. For such a system, a representation other than a simple image may be more appropriate.

There is a fair amount of empirical data on the geometry of the representation of visual space in V1 and this shows that the representation is far from a simple image. Important for the proposal here, there is also evidence that input correlations are implicated in the formation of this representation in V1. A theory of the structure of this representation should not only taken into account the geometry of the representation in V1, but what this representation is used for, and how it is formed.

After reviewing the data on the structure of one particular representation of space in cortex, that in V1, previous mathematical characterisations of this map be summarised. A number of models of the development have been proposed, and although they do

result in topographic maps, it is argued that they both fail to cope with the distributed activity present in the input (in an image, there is activity throughout the image). They also fail to deal with the fact that the function mapping from the retinotopic world to the the cortex is not the identity map.

Inspired by observations that correlated input activity is implicated in the formation of topographic maps, and that in the resulting map the fovea is vastly over represented, a model of cortical map formation is proposed. The resulting map proposed by this theory is determined by the input correlations, and by measuring these correlations, the exact form of the resulting map can be found. The structure of this resulting map is then compared against the form of the map of V1 in cortex as inferred physiologically. Finally the reason the model matches the data so well is discussed in term of the need for a map that is optimised not for statically viewing the world, but actively sampling from it.

5.2 Previous work on topographic maps and their formation

5.2.1 The geometry of Striate cortex

The visual area V1 has a very noticeable topographic organisation. Cells that are close together in cortical space have receptive fields that represent measurements that are close together in retinotopic space. This mapping seems in maintains topographic relationships, but is not the identity mapping: the geometry of the input is not the same as the geometry of retinotopic space. This leads to the question, what is the form of the function that maps the geometry of input space to that in cortex? Although very time consuming experimentally requiring the recording from large numbers of cells, the identification of their receptive fields, and anatomically locating the cells in V1, a number of studies have both experimentally measured this relationship, and attempted mathematical characterisations of this mapping.

Daniel and Whitteridge (1961) introduced the concept of the cortical magnification factor (M), the number of millimeters traversed in the cortex divided by the number of degrees of visual angle in the visual field traversed by moving this distance in cortex. M was shown to depend on the degree of eccentricity (E) getting smaller as eccentricity increases (Daniel & Whitteridge, 1961; Hubel & Weisel, 1974). This was measured and it was found that the magnification factor scaled approximately with 1/E and therefore eccentricity in the cortex corresponded to a log function of eccentricity in

world coordinates.

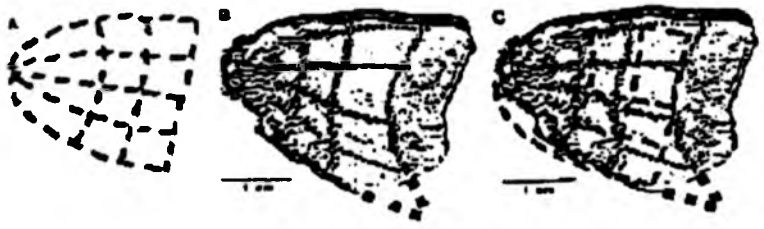

Figure 5.1: The theoretical map of three different circles (eccentricities 1°, 2.4°, and 5.66°) as modeled by the complex function log(E + 0.3) proposed by Schwartz (1985). **A** The structure of the complex function log(E+ 0.3) **B** The actual representation of these equal eccentric circles in Macaque V1 found by Tootle et al (1982). **C** The fit of the model (the function log(E+0.3)) to the data. Note that the fit is approximately correct, but the model overestimates the represented distance in the horizontal meridian (the center), and underestimates the distance in the vertical meridian (the upper and lower part of the visual field). The simple complex logarithm map cannot account for these differences, but the correlational model proposed here can. (from Schwartz, 1985).

Tootell et al (1982) measured the representation of the retinotopic distance in V1, for the representation of the central part of the visual field (1° to 5.66°). They found an approximate logarithmic transform from retinotopic coordinates to cortical coordinates, but found that the represented distance along the vertical meridian was represented as approximately 1.25 times longer than that of the horizontal one. They also argued that the representation was non-conformal[1], that is, angles were not preserved in the cortex as would be expected if the expansion was purely a complex logarithm as proposed by Schwartz (1980) . This was a source of controversy. Schwartz argued that distortions from conformal were caused by stretching of the samples (Schwartz, 1985), whilst Tootell (1985) claimed that the representation was non-conformal, and argued for an explanation in terms of the need to interlace ocular dominance columns. Later both accepted that the logarithmic approximation was approximately correct, but that the structure of V1 probably had considerable shear and hence was non-conformal (Landau & Schwartz, 1992).

[1]Of functions that map one geometry to another, one that has proved useful in other areas is the class of conformal mappings. A conformal mapping is one where, even although the distances in the two representations are changed, local angles are preserved. This would mean that a right angle in the world would map locally to right angle in the representation. One particular class of conformal maps is the logarithmic maps that go from a a point (with polar coordinates r,θ) in the world, to a point $m \log(r + c)$, theta. This was used by Schwartz to model the transform from retinotopic coordinates to cortical coordinates.

Van Essen et al. (1984) also measured the representational geometry of V1, but using a different technique measured it for a much larger part of the visual field (2.5° to 80°). They again found an approximate logarithmic transform, but pointed out a number of anisotropies in the representation. Again the representation of an angle in the vertical direction extended for an increased distance, more cortex was also found to be devoted to lower than upper visual fields, and more cortical area assigned to the ±45° around horizontal, as opposed to the vertical meridians (see Figure 5.2).

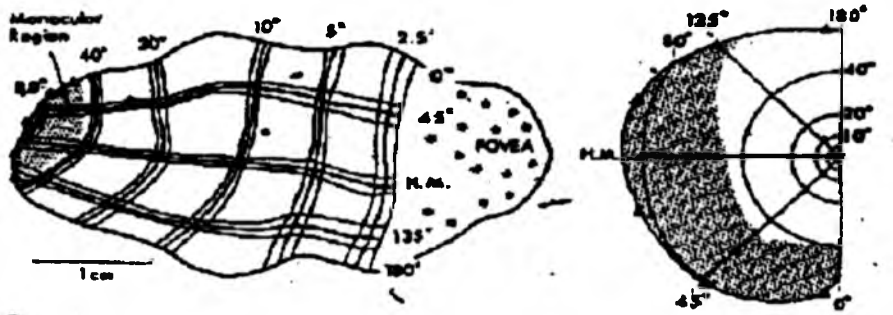

Figure 5.2: The lines of constant eccentricity as found experimentally by Van Essen et al (1984). Note that although eccentricities increase by a multiple of two, the distance between contours is approximately constant indicating an approximate logarithmic mapping between eccentricity and represented distance. Note also that the horizontal meridian is represented as closer to the fovea than the vertical meridians.

Their study was extended to a number of different monkeys and considerable variability was found between representations. This could not be accounted for purely by the different sizes of V1 between the macaque monkeys used: the "error bars" of this variability were found to be approximately constant in cortical space and therefore approximately proportional to the eccentricity in real space (assuming an approximate complex logarithm mapping).

Schwartz (1985) took the data from the Tootle study, and using the logarithmic approximation, explicitly fitted a function of the form: represented distance $\propto$ log(eccentricity + c). The free variable c was found empirically to be 0.3 for the macaque monkey. Mallot (1985) proposed an alternative mapping function, a complex power law with an exponent of p = 0.43 for V1. This was argued to have more generality than the complex logarithm transform usually proposed (Mallot & Giannakopoulos, 1990).

5.2.2 The role of correlated activity in map development

The initial connection from the lateral geniculate nucleus (LGN) to V1 possesses only crude topographic ordering, and the initial representation is not activity dependent (Meister *et al.*, 1991), but depends on a chemical gradient system. The subsequent refinement of the representation requires correlated input activity (Shatz & Stryker, 1978). Blocking input activity from the retina using tetrodotoxin blocks the refinement of the representation. Inducing correlated activity via electrodes changes the form of the representation (Fawcett & O'Leary, 1985).

Waves of correlated activity have been observed sweeping across the retina at about 100 microns per second then followed by a period of inactivity. These have been implicated in the formation of topographic maps pre-natally (Shatz, 1992). Continuous modification of the representation has also been observed in adult cats. By changing the input statistics by suppressing the activity presented to a small part of the input, the receptive field sizes can be made to reversibly expand in the order of ten minutes (Pettet & Gilbert, 1992). Selective lesioning of the retina using lasers can also change the topography within V1. Over a period of 2 months, the cortical area that is used to represent the lesioned area is reorganised to represent unlesioned areas (Gilbert, 1992; Gilbert & Torsten, 1992).

5.2.3 Models of the activity based formation of the maps

Andrews (1964) proposed that various statistical redundancies within the visual world could be used to learn the metric for the psychological representation of space. Two of these proposed regularities were that contours are equally spaced throughout the visual field, and these contours are on average straight. These were proposed to be used by the cortex to learn the correct metric for a system which was in initial correct topological correspondence with the world (the representation had no twists, kinks or tears), but the distances were not correctly represented. By changing the representation until the measured statistics had these characteristics, the world could be used to calibrate the representation.

It was hypothesized that the statistics would vary across the visual field. Fixations would often be on edges (the center of the visual field), edges would be sparse in the upper field (because there would be few objects), and an intermediate number of edges would occur in the bottom of the visual field. If these distorted statistics were used, this would induce distortion in the representation causing compression in the upper and

lower visual field. This was related to the psychological evidence that the representation of space is supposedly approximately hyperbolic.

Another paradigm has concentrated on computational models of the formation of topographic maps. Prestige and Willshaw (1975) produced a network that formed topographic maps based on two proposed chemical marker systems. Willshaw and von der Malsburg (1976) implemented a network that used spatially correlated activity in the form of dipoles to form a topographic map, given a network with a small degree of initial order. Von der Malsburg and Willshaw (von der Malsburg & Willshaw, 1977: Willshaw & von der Malsburg, 1979) implemented a model that used induction of chemical markers from the retina, and Cowan (Whitelaw & Cowan, 1981: Cowan & Friedman, 1991) proposed a model that combined both spatially correlated activity and a chemical marker system.

Only one model however, has attempted to account for the distorted representation of space. Amari and later Zhang (Amari, 1983; Zhang, 1991) performed a mathematical analysis on a self organising map. Assuming that the input to the net was spatially localised (limited spatial extent of auto–correlation), the magnification factor and receptive field sizes of cells in different parts of the map were calculated. Under the conditions studied it was shown that areas of the representation that received more input would be represented in more detail, and that predictions on the relationship between the magnification factor and the size of the receptive fields match those found in cortex.

Despite this, the analysis only had a limited applicability. Only localised groups of input and output activity could be dealt with. The map expansion and magnification factor predictions came from assuming that simply more "events" happened in the fovea than the periphery and no mechanism for this was proposed.

In summary, the cortical area V1 is in topographic correspondence to the world. This mapping can at least approximately captured in terms of a logarithmic transformation between retinal eccentricity and cortical eccentricity. Despite this, there are a number of characteristics that do not fit this characterisation: the different representation of the vertical and horizontal, the upper and lower visual field, and the parameters used to fit the map are entirely arbitrary. Correlated input activity has been implicated in the fine tuning of these maps, and there exist a number of models that given correlated input activity, can form maps in topographic correspondence with the world. Despite this, all these models usually assume only local input activity (natural images have activity throughout the image), and although they can produce topographic maps, the

maps produced are usually identity functions not the approximate logarithmic maps as found in cortex. What is required is a theory that can use the input to recalibrate its representation and generates not an identity mapping, but one that is approximately logarithmic. This mapping should not be precisely logarithmic, but deviate from the exact logarithmic function in the way that has been observed empirically. The parameters of this function should also not be arbitrary, but in some form determined by the nature of the model. Lastly this model should be plausibly applied not just to isolated "events" as most topographic mapping models assume, but an input where all the input is active at any one time. The next section details a model that attempts to answer these questions.

5.3 A model of the relationship between cortical representation and input correlation

As well as the obvious topographic organisation of the representation, it is obvious from the experimental data that the representation of the fovea is privileged. This location is functionally more important (if something is of interest, it is fixated). It is also physiologically more important, with the amount of cortex devoted to representing the fovea enormously greater than that devoted to representing eccentric locations. Therefore any theory that ignores the privileged nature of the representation of the fovea cannot possibly account for the function mapping from the world to V1 because experimentally, the mapping is so different at the fovea than at eccentric locations.

Another important factor to be taken into account is the implication of input correlation in the formation of spatial representations. As seen in the previous chapter, an assumption that distances in psychologically inferred spatial representations are calibrated using input correlations produces reasonable predictions. This assumption of a correlation based metric has also has been seen to be inherent in the operation of many models of cortical map formations (Goodhill, 1992), and although these models fail to account for the geometry of the representation, they provide reasonable (and testable) accounts of other mapping phenomena such as the development of ocular dominance.

As argued previously, models simply using input correlation can account for the structure of psychological space, but because no privileged status is given to the representation to the fovea (or any location), it cannot possibly be used to understand the structure of cortical maps. This suggests a very simple modification to the correlation based metric theory proposed in the previous chapter. Rather than using correlation

to calibrate the representation of the distance between any two arbitrary points, we propose that the representation of fovea is privileged, and that correlation is used to calibrate the represented distance of a given measurement from the fovea.

To be more specific, if N_{fovea} is a neuron representing a measurement at the fovea, and $N_{OtherLocation}$ is a neuron representing another location, then in V1 in the cortex, the neuron $N_{OtherLocation}$ is represented a distance from the fovea proportional to one minus the correlation of the measurements made at these two locations:

$$E_k \propto 1 - correlation(M_k, M_f) \tag{5.1}$$

where E_k is the eccentricity of the neuron k in cortical space, and M_k are the set of measurements made by this location k, and M_f is the set of measurements made at the fovea. If a scheme such as this was used, a measurements made by a neuron that was highly correlated with measurements made at the fovea, then in cortical space the neuron making these measurement is represented as very close to the fovea. If alternatively the measurements made by a neuron have a very low correlation with measurements made of the fovea, then this neuron will be represented as far from the fovea. This therefore specifies a relationship between input correlation and the function relating eccentricity in the world to eccentricity in a representation and therefore a potential signal that could be used to calibrate a visual representation.

This also constitutes at least a component of a model of the structure of the representation in V1. Although there is no algorithm to implement this mapping[2], for a given input correlation structure, we can infer the form of the function relating eccentricity in the world to represented eccentricity in cortex. This model has the virtue of simplicity. It also proposes a privileged status for the representation of the fovea and therefore has the potential to account for the observed maps. It works on the input correlations, and therefore is commensurate both with the psychological phenomena presented in the previous chapter, and the implied role of input activity in cortical map formation. Lastly it is inherently very testable. Given the form of input correlations as found in the world, the theory predicts the form of the function mapping from world to cortical space. This function can then be compared to the empirically observed map and this is done in the next section.

[2]One is currently being developed using a probabilistic framework, where priors are expressed on the inter "hidden" unit correlations.

5.4 Experiment 12: the spatial correlations of the world revisited

As in the previous chapter, the predictions of the model depend crucially on the spatial correlations in the world. This being so, we again sample from a set of natural images, but this time the samples taken are larger, the exact set of images is different, and the ways we view the resulting correlation structure is different.

5.4.1 The images

The images again consisted of a collection of pictures of natural scenes intended to roughly mirror natural visual experience. The pictures were taken on a 35mm camera with a 50mm lens (unless otherwise specified). The photographs were then digitised using a Hewlett Packard Scanjet-plus at the 75 dpi setting. The central 256 by 256 pixel region was then sampled and the gray levels normalised so that the lowest value was 1 and the largest was 256. No attempt was made to account for non linearities in the processing. This process was used to create the following 87 images, many of which were used in the experiments in the previous chapter:

- 15 images taken from the office environment (people, car parks, furniture, corridors etc).
- 15 images from a flat flood plain area (river scenes, fields, and country roads).
- 14 images taken within a city center (crowds, houses, shopping centers.).
- 14 images of plants and vegetation (heather, roses, trees).
- 12 scenes from hills and mountainous countryside.
- 10 images of peoples faces (male and female).
- 4 pictures of animals in natural surroundings (a bird, sheep, cats).
- 1 man made sculpture.
- 1 image of a page of text.
- 1 view out to sea.

Equivalent angles were calculated as in the previous chapter by taking a picture of a ruler at a distance of 1 meter, then digitising it with the other images. From this, it was calculated that $1^\circ \approx 10.9$ pixels.

5.4.2 Finding the spatial correlation

The same technique was used to find the correlations as described in section 4.4.1. This consisted of taking large number of samples from the image set, and then using these to calculate the correlation of all points with the central pixel across all the samples. The main difference here is that the sample size chosen was not 64x64, but 128x128. By using this larger window size, not only can we estimate the correlations over a longer distance, but this naturally constrains the fixation point (the center of the sample) to the center half of the image[3]. This larger sample size therefore prevents samples of the sky, the bottom corners of the image, and other locations that are probably less representative of the image statistics as sampled by monkeys (and man).

5.4.3 The results: the correlation structure viewed in a different way

This sampling and correlation calculation produces an estimate of the correlation of eccentric locations with the central location in a sample. This is similar to the correlations found in the previous chapter except for the larger samples size. The physiological observations are expressed in terms of contours of constant retinotopic eccentricity plotted in cortical coordinates. To facilitate comparison with this experimental data, the correlations as measured can be used to plot what these contours would look like according to the model (equation 5.1) and the measured correlation in the images. This is shown in Figure 5.3. Here the contours for seven different eccentricities are plotted as they would be if the cortex was organised in the manner proposed by the model (see figure legend).

5.4.4 What can be see from the correlations?

Even without quantitative measurements, two aspects of the measured correlations, and hence the inferred mapping between retinotopic and cortical space, qualitatively match the experimentally measured maps. In Figure 5.3, each circle in retinotopic space has a radius that is a constant multiple larger than the last. The circles plotted in "cortical" space are not multiples further apart, but at least approximately constant shifts and this corresponds approximately to the experimentally observed logarithmic mapping. The circles also do not map to straight lines in "cortical" space but extend

[3]The central pixel of the sample or fixation point can only exist in the region x = 64 - 192, y= 64 - 192, and therefore samples are only taken from the center of the image

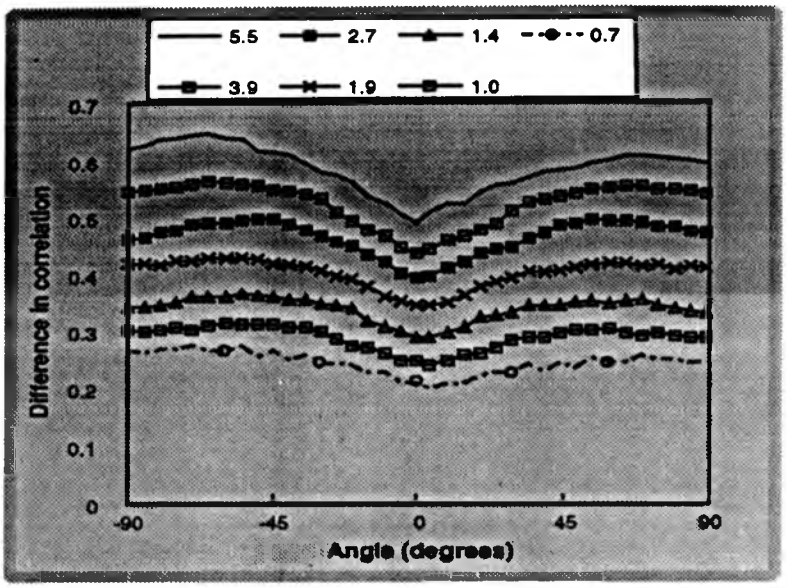

Figure 5.3: The correlations as measured in the 87 image set but displayed in a manner that aids comparison of the model predictions and the empirically observed map in V1 of the Macaque. Along the x axis is plotted the angle from the center of the sample with 0 degrees being the horizontal direction (east by compass direction), +90 degrees being directly downwards (south), and -90 degrees being north. Along the y axis is plotted the value of 1-the correlation. The seven different lines then correspond to seven circles in retinotopic space as they would be seen in the simulated "cortex", the first circle has a radius of 0.7 degrees and the subsequent circles are each root 2 times greater in radius. If the correlations measured in the world were isotropic, then these circles in retinotopic space would map to straight lines in this graph, but since the correlation varies with angle, they are in fact far from straight lines. The correlation is plotted in this way because if it is true that eccentricity in cortex is proportional to correlation, these lines will correspond to the contours observed in cortex. This means that comparison with experimental data is simplified.

further in the vertical direction than in the horizontal direction and this was observed in both the Van Essen and Tootle studies. In the next sections, it is attempted to turn these qualitative observations into quantitative results by comparing parameters of the inferred map with those measured in the physiological studies.

5.5 Experiment 13: the parametric form of the function mapping from retinotopic to cortical space in both model and cortex

The relationship between eccentricity from the fovea in the world, and the represented distance from the fovea in the V1 has been explicitly modeled by two functions. Firstly, it is widely agreed that the function is approximately logarithmic (Schwartz, 1985; Schwartz, 1980; Tootell *et al.*, 1985). This suggests using a parametric logarithmic functions to fit the macaque data and this was done by Schwartz using a function of of the form:

$$R \propto \log(E + c) \tag{5.2}$$

where R is the distance from fovea in cortex, E is the eccentricity in retinotopic coordinates, and c is an experimentally fitted constant. For the Macaque, this constant was found to be 0.3 (Schwartz, 1985). Mallot et al (1990,p247) argue that using power law functions of the form $R \approx E^p$ is more appropriate (fitting p). For the Macaque, Mallot et al found the best fitting value of r to be 0.43. Given the estimate of the spatial correlations in the world, it is possible to find the best fitting parameters for the map implied by the model and this was done.

5.5.1 Finding the best fitting parametric fit to the correlations

To find the best fitting parameters to the function mapping retinal eccentricity to cortical eccentricity, the correlations in the vertical and horizontal directions were found. In the previous chapter, we used the observed variability in the correlations to weight this functional fitting, but since this was not done in either the Schwartz or the Mallot study, this was not done here. Instead the flexible function fitting capabilities of the KaleidaGraph package were used to find a minimum squared error fit of functions of the form:

$$R \propto 1 - \phi = m1 \log(E + c) + m2 \tag{5.3}$$

and

$$R \propto 1\rho = m1E^p + m2 \tag{5.4}$$

with m1 and m2 being arbitrary parameters, and c and p being the parameters found in the previous studies by Schwartz and Mallot. The gradient descent method

Fit	m1	m2	c	χ^2
horizontal	0.46	0.25	0.32	0.0004
vertical	0.35	0.22	0.25	0.002

Table 5.1: The logarithmic function fit to a vertical and horizontal slice through the correlations. This is an extremely good fit to the empirical data where c is estimated to be 0.3, especially as the log fit to the empirical data is much better to the horizontal meridian (see Figure 1 in Schwartz (1985)).

used by KaleidaGraph was started at a number of different start locations insuring local minima were not a problem. The best fitting parameters together with the machine generated *chi*^{2} were then recorded.

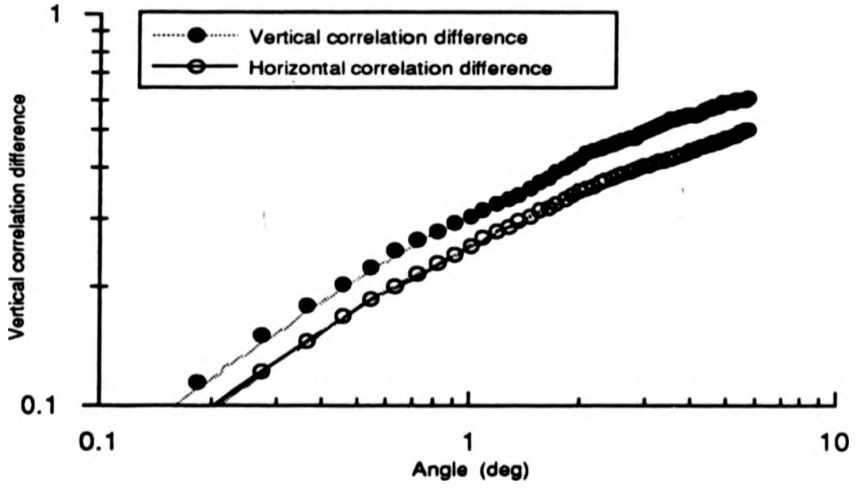

Figure 5.4: A plot of correlation versus eccentricity for the vertical and horizontal direction. The line through the data is a log(E +c) fit to the correlations. This line faithfully captures the trend in the data with c = 0.32 for the horizontal slice. Schwartz (1985) fitted a logarithmic function to the relationship between eccentricity and cortical distance and found a Figure of c= 0.3 for macaque. This fit together with a power law fit to the model (see text), are very close to those found empirically by Schwartz (1985) and Mallot (1990).

The result of logarithmic law fit as proposed by Schwartz is shown in Figure 5.4 with the parameters being given in Table 5.1. As can be seen the fit is excellent with the fit in the horizontal direction, giving a value of *chi*^{2} of 0.0004, and in the vertical direction 0.002. The best fitting value of c was found to be 0.32 for the horizontal

direction and this is extremely close to the value inferred from physiology by Schwartz of 0.3. For the logarithmic function fit, the model and experimental results are in very close agreement.

A power law fit of the form E^p was also made to compare to the fit for the macaque given by Mallot (1990). Again the fit was very good to the data with the best estimate of p being 0.42 for the horizontal and p = 0.46 for the vertical (compared to p=0.43 found empirically by Mallot et al (1990)). In conclusion, the form of the function inferred from the model (approximately logarithmic), fits that found experimentally very well. Not only this, but parameters that were completely arbitrary under the previous models are also fitted reasonably and certainly within the realms of experimental error.

5.5.2 The different representation of the two meridians in model and experiment

One aspect of the experimentally inferred maps that has caused a lot of controversy is that in the maps measured in cortex, the representation for vertical and horizontal is different (see Figures 5.1 and 5.2). This on the surface seems awkward for the interpretation given by Schwartz that the maps are conformal as such shears in a map, as pointed out by Tootel et al (1982) is not conformal. To explain such effects, it is proposed that this is caused by the ocular dominance stripes being interlaced in this direction.

This slightly ad hoc explanation is not required in the correlation calibration framework. As can be readily seen in figure 5.3, the map as generated from correlation would also represent the two meridians differently. Of interest then is whether the ratio of represented length in the vertical to horizontal direction generated by the model is similar to that found in cortex. Schwartz (1985) states that the ratio of the represented length of the vertical meridian is longer than the horizontal meridian by a ratio of 1.25. Using the correlations shown in Figure 5.3, the ratio of the represented distance in the vertical and horizontal directions was calculated for circles of 5 eccentricities. The results are given in table 5.2. As can be seen, the ratio that would be observed (if the image set used is representative) from the model is roughly 1.23. This again is very close to that found empirically of 1.25. Note also, that although it is a small effect, the circles in Figure 5.3 extend further in the upward direction (-90), than in the downward direction (+90). This again corresponds to an anisotropy observed by Van Essen et al in the experimental data where they observed that the representation of upper meridian is represented as further away than the lower meridian.

Eccentricity	Represented distance Vertically	Represented distance Horizontally	Ratio
42.5°	0.556	0.444	1.25
30°	0.494	0.400	1.24
21°	0.420	0.352	1.20
15°	0.36	0.29	1.24
10°	0.31	0.251	1.24
Average ratio			1.23

Table 5.2: The ratio of the represented distance in the vertical and horizontal direction as predicted by the model. The average value found was 1.23, this compares very well with the empirically measured value of 1.25 (Schwartz1985).

5.5.3 The within species variability of the representation

The last aspect of the representation that can be looked at is the variability between different macaque monkeys. The initial data was based on the intensive study of a single monkey. In addition, a number of other monkeys were studied in less detail. Although each monkey showed little variation between representations of the two hemispheres, there was large variation between monkeys (see Figure 5.5). The estimated location of the represented points for 2.5°, 5°, 10°, 20°, and 40° eccentricity was plotted. The spread of these was approximately constant (see Figure 5.5). This will cause problems for networks where the topographic mapping is induced by the local interaction. In such a scheme it would be expected that the variability would increase through the representation (That is, it would be cumulative). For the system proposed where the measure is not local, the form of variability is simply explained as a constant variation in the estimate of the correlation. As seen in chapter 4, this leads to a variability in real space proportional to the distance, and therefore a constant error in the logarithm of the distance.

5.6 Discussion

5.6.1 V1 as a representation optimised for active sampling from the world

The simplest way mapping between the world and cortical coordinates would be an identity mapping, but very obviously this is not the case for the mapping found in

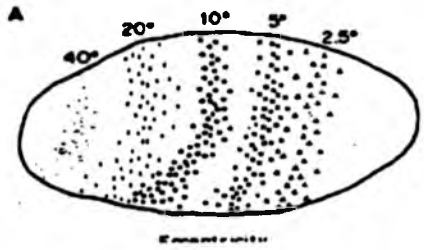

Figure 5.5: The variability in the representation of different eccentricities as found by Van Essen (1985). Note that the scatter for each eccentricity is approximately equal for all eccentricities.

V1. Why? As described earlier, Schwartz proposed that the mapping was a conformal complex logarithmic mapping and argued against the reliability of evidence against this proposal. The reason for this was his elegant proposal that a complex logarithmic mapping gives good invariance properties. If a given object was fixated in the center, then if the object changed size (because for instance it moved nearer of further away), then in cortical representation this would correspond to a simple shift in cortical co-ordinates. This converting a desired invariance to change in size to a simple cortical transformation was proposed to make invariant recognition much simpler.

Whilst it is true that the complex logarithmic mapping will map size changes of centrally fixated objects to shifts, as an explanation of the form of representation, this has many problems. Firstly, if the object is not quite centrally fixated, then a size transformation causes a very different representation to result. The map as observed is also not pure complex logarithmic, apart from the different representations of the vertical and horizontal meridians, the map in V1 is different for the upper and lower visual field and it is very difficult to interpret this in terms of Schwartz's proposal. Lastly, it is far from clear why recognition of a shifted representation is any easier than the recognition of an expanded object. If this is not easier, then the whole justification of Schwartz's proposal seems suspect.

The observation that cortical representation scales with input correlation suggests a much simpler interpretation. In an animal that actively fixates its environment, the fixated point is of far more interest than eccentric locations. Despite this, the identification of a fixated location is often aided by local context and information about more eccentric measurements made form the fixated object. When viewing a single object, rather than an image of the whole scene, what would often be more useful would be a representation of just this object, excluding all other locations. Unfortunately ob-

jects vary greatly in size, and a representation that was appropriate for small objects would be useless for fixating larger objects (or at least require far more fixations). As a compromise, we could devote representational power to eccentric locations that was proportional to the average chance that they were relevant to the identification of the fixated object. Although not always true, a single object will often have similar characteristics, and therefore the correlation of eccentric characteristics gives an approximate estimate of, on average, how relevant a measurement made eccentrically is relevant to the identification of the centrally fixated object. It is therefore proposed that the representation in V1 is therefore organised approximately logarithmic simply because we want to devote more representational power to the point of fixation, but still represent eccentric locations as they often come from the same object and therefore aid identification of this object. The correlation of these two measurements gives some indication of how often these two measurements are made from the same "object".

5.6.2 Conclusion

The preceding chapter included an investigation of the possibility that correlation with the fovea is used to calibrate the early cortical representations of space. The following were proposed:

- In a system that actively foveates the world, what is of interest and therefore represented are measurements relating to the fovea.
- One way to estimate the average relevance of a measurement in the periphery to measurements for the fovea is to find the correlation between the two measurements.
- The relevance as measured by correlation can be used as a metric for organising the representation. Highly relevant measurements are represented close to the fovea while irrelevant ones are represented as far away.
- The correlations were measured from a large number of images, and the form of the representation for these scenes is calculated given that the proposed model is true.
- A qualitative comparison of the model to the functional geometry of the macaque striate cortex shows a very good match. Most importantly, not only the form of the relationship between eccentricity and represented distance, but also the

parameters match. The ratio of horizontal to vertical meridians also matches that found in cortex.

Previous proposals for the formation of topographic maps have been of two types. Some concentrate on the development of the map, and have avoided the problem that the map has a different topology from that of the input. Others such as Schwartz (1980) have concentrated on the computational advantages given by a complex logarithmic representation, whilst not concentrating on either the non-conformal nature of the representation, or how the representation is formed. The model presented here has many of the advantages of both approaches in that, the mechanism is sufficiently similar to those previously proposed so as to possibly have implementations using simple biological mechanism. It also gives a reason for the form of the representation, not that scale relationships are treated as shifts, but that when sampling from the world, what is of interest is what is foveated. This simple assumption gives accurate quantitative predictions of the form of representation, providing evidence in favour of this interpretation.

Chapter 6

The estimation of posterior probability of interpretations using phase coherence in neural networks

Summary

As well as inter-layer connections, the visual cortex has a large number of intra-layer connections. It is proposed that the purpose of these connections is to represent a probabilistic model of previous experience, with the Boltzmann machine learning algorithm being one method of constructing such a model. A model of this form has a number of potential uses. It can be used to find out how probable a particular interpretation of an input was in the light of previous experience. It can be also used to find aspects of the signal that do not correspond to the model, and if the input as a whole was probable, then the model can be used to "correct" these exceptions.

In a model such as the Boltzmann machine, one quantity of great importance is the energy of a given state (combinations of features). This quantity is proportional to the log of the probability of the state, and is therefore useful for conveying a states probability to later levels. It can also be used whilst searching to determine whether a good state has been found yet, and to assess whether the model is for this particular input.

Here it is argued that if a network is constructed out of biologically realistic spiking neurons, which can communicate both via a firing rate and by using the timing of the spikes, then there is a simple relationship between the coherence in the spiking times and the energy of the state. Computational experiments are presented showing a regular relationship between spike train

coherence and the energy of the state of the network. This result means that an estimate of the probability of a state of a network can be assessed within the second spike transmitted by a neuron (in the case studied here). This it is argued give an alternative explanation to the observed phase locking behavior observed in cortex(Singer, 1990). The degree of phase locking between units is not used to link features, but used to communicate the probability of an interpretation of an input to later levels. In conclusion it is argued that given an indirect estimate of the energy of a network state, this can be used to communicate the probability of an interpretation, control the temperature if a simulated annealing type search method is used, and moderate the application of the model to the data dependent on how appropriate the prior is to the data.

6.1 Introduction

The most widespread framework for understanding the operation of recurrently connected neural networks is that of the Hopfield network. Implicitly the role of such networks is that of auto-association. In auto-association, previous knowledge of the input is represented as a number of memories and the role of such a network is, given an input, to find the previous memory that is closest to the given input. Much of the within area connectivity of the brain is recurrent: a given neuron is connected to other neurons, which are in turn connected back to the original neuron. The architectural similarities of this to the Hopfield network could possibly suggest the role of auto-association for these connections (as for instance in (Amit, 1989)).

As well as the auto-associative model prevalent in research on recurrent neural networks, there is another framework that uses a recurrent architecture. Here the role is not the storing of memories, but that of modeling the probability distribution of the input. This model can be used to regularise an input: if given an image which is ambiguous in its interpretation, the model can be used to choose the most probable amongst the many interpretations. Whilst understanding the computation of areas such as the hippocampus, modeling in terms of the storage of memories may be appropriate, it is argued here that in low-level vision, forming and using a probabilistic model of the the input is much more relevant operation than the storing of a limited number of memories (whatever these would be in the case of low-level vision).

Such probabilistic models can be implemented using a recurrent architecture of the

same form as the Hopfield model such as the Boltzmann machine, but this is not without problems. In a system with many potential sources of information as to the state of the world, we have have not only to convey the most probable state of the world, but also need to label the particular interpretation as to how probable it is. The Boltzmann machine provides a mechanism for doing so (it signals all possible interpretations with the probability that a given state is signaled being directly proportional to the probability of that state given the model), but this method is unfeasibly slow as a metaphor for the operation of a biological system. Also, if the model is to be used to "correct" an input, we first must in some way check that there is previous experience appropriate to such situations. If the input is very improbable under the model, then this indicates that it is unlike the data that was used in forming the model, and therefore will not be of much use in correcting the input. One method of avoiding such problems is to only use the model to correct an input if the input was probable under the model.

Lastly, unlike in the traditional auto-associative role where the the memories are stored in terms of local minima of the energy function (see later), in the probability distribution modeling framework of the Boltzmann machine, local minima are not of interest because we would like to signal the most probably interpretation and not just a local minima. Therefore some method to avoid the worst minima is required. This avoiding of most local minima is difficult even in computer vision applications of such models, and for such models to be plausible as biological metaphors, some method has to be found to avoid the worst of the local minima.

The next two chapters address potential solutions to these problems. The main thesis of this chapter is that partial solutions for the above problems can be found if at any given state, we have some indirect estimate of the probability of the current state. It is proposed that if instead of the simple probabilistically firing neurons of the standard Hopfield and Boltzmann machine networks, we use more biologically accurate neuronal approximations, then computational experiments can show that in the configurations studied, the probability of the current network state is directly represented in the entropy of the spike emission times (a measure of spike time coherence or degree of phase locking). A measure of the probability of a given state in this form can then be used to signal this probability in a usable form to later levels, moderate the application of the model, and possibly control the application of the model.

The rest of this chapter has the following structure: the Hopfield network is described together with reasons why this is not a practical framework for modeling low-level vision. The alternative framework of forming and using a probabilistic model is

then discussed together with its advantages. The Boltzmann machine, a neural network that forms probabilistic models using an architecture almost identical to that of the Hopfield is described. Some of the the problems in using such as system when there is not enough time to observe the outputs over a long period of time (as would be the case in a biological system), are discussed. Computational experiments are then performed showing how the energy of a given state (described later), and hence its relative probability, can be found in simple cases by just measuring the degree of coherence in the spike trains between the different neurons. The discussion then argues that a measurement in this form is very useful in a biological system to communicate the confidence in a solution to later levels, to control the "temperature" used when performing search, and in choosing how strongly to apply the model to the data.

6.2 Hopfield networks

Hopfield's finding that a network of simple idealised neurons, symmetrically connected can operate as an auto-associative memory (Hopfield, 1982) has been tremendously influential. This network consists of a large number of binary threshold units each connected both to the outside world and to all other units via real valued weights. The operation of this system then just consists of randomly choosing a unit and recalculating its input:

$$I_i = \sum_{j \neq i} w_{ij} S_j + \eta_i - \theta_i \tag{6.1}$$

Where S_i is the output state of unit i, $w_{ij} = w_{ji}$ are the symmetrical weights between units, θ is the threshold of the unit, and η is the external input from the outside world. The output is then updated either deterministically:

if $I_i > 0$ then $S_i = 1$ else $S_i = 0$

or stochastically:

$$P(S_i = 1) = 1 - P(S_i = -1 or 0) = \frac{1}{(1 + \exp -I_i/T)} \tag{6.2}$$

This is a potentially interesting procedure because, using methods from statistical physics, for any state of the network (combination of individual units states), an energy (E) can be associated with that state:

$$E = -\frac{1}{2} \sum_i \sum_{j \neq i} w_{ij} S_i S_j + \sum_i \theta_i S_i - \sum_i \eta S_i \tag{6.3}$$

For the deterministic network (one where the output is determined purely on whether the input sum is above 0), the dynamics always leave the energy the same or reduce it. For the stochastic network, the probability of being in any state at equilibrium is proportional to $\exp -E/T$. This means that if the weights, thresholds, and inputs from the outside world can be set to values that make E an important quantity for the system to reduce, this architecture can be used to combine information from the outside world and the model in a principled way.

In the Hopfield network, the energy is used in the following way. We have some "memories" which consist of a number k, of binary valued vectors M (either having values of +1,-1 or 1,0) of length N. If the weights are then set to $< M_i, M_j >$ where the expectation is over all the memories, then if the network size (N) is large, the patterns are nearly orthogonal, and the number of memories (k) is small, then these memories become local minima in the energy function of the network and hence stable points in the network dynamics (for a very much more detailed account see (Amit, 1989)). This means that if the state of the the network is initialised with an input from the world, then the dynamics will change the system until it is in one of these memory states. Often this state will also be the nearest memory state to the input (but especially in small networks, this is certainly not always the case). This means that the network can act as an auto–associative memory.

This ability to assign an energy function to any state of the net has allowed the techniques of statistical mechanics to be applied. This has made the Hopfield auto-associative network one of the best understood analytically. This analytic tractability has allowed detailed studies of the storage capacity (Gardner, 1987), the effects of noise, and modifications that extend the biological plausibility of these admittedly simple models (Amit, 1989). Despite this, the practical applications of such models have been extremely limited, as has its use as a biological model.

This has primarily been for two different reasons. Firstly, as an auto-associative memory, the Hopfield prescription is not very good. The storage capacity of these networks is very limited at about 0.14 times the number of units in a fully connected system, and this is only true when the network size is very big. When run at anything other than an optimal temperature [1], a large number of "spurious" attractors are present

[1] In a system run with no noise (called zero temperature), as well as the memories, a number of non-memories become stable states of the network. If run at high noise (or high temperature), many of the memories cease to be stable. The finding of the correct level of noise in a system is difficult, especially as experience of the author with non infinite sized nets indicates that often the "optimal" temperature found empirically is different from that given analytically.

as the memories. These correspond to combinations of various memories. Although these are potentially useful, in the auto–associative framework these can only be viewed as spurious ((Amit, 1989) p 199). Even this limited storage capacity relies on the memories being completely uncorrelated. Very few interesting data sets consist of data with no internal correlational structure, and even such simple regularities as the average chance of a bit being true not being 50% can cause problems.

The second and more fundamental problem is that for the majority of problems, including that of modeling low-level vision, auto-association is not a useful operation. Expressing all previous knowledge of the world in terms of a limited number of "memories", then replacing any signal encountered in the world with one of them is not a good metaphor for perceptual processing. Given an input unlike any previously encountered, the resulting output is going to be an arbitrary previous memory, and if the input is unlike anything in previous experienced, then the system will convey essentially no information about the input.

Again we have problems when we encounter an input similar to two different stored patterns. The most sensible approach would seem to be to use information on the similarities to both patterns in order to decide what to do with the input. This is not the behavior prescribed by the auto–associative paradigm. In this we find the closest pattern and just signal this: all knowledge gained by its similarity to other stored memories is lost. If the input is equally close to two patterns, information from both should be used in restoring the input, and signaling one of the memories arbitrarily is not sensible. The relationship between individual features and more local structure is important, not just the actual memories that we find this structure from.

6.2.1 Regularisation theory.

An alternative method of using knowledge of previous inputs that has proved of far more practical value is statistical regularisation. In many perceptual problems (eg deconvolution, shape from shading, shape from texture), the problem is ill posed: the information in any particular input is insufficient to determine a solution uniquely . To solve these problems additional prior information about the world is used and this prior information or preference for some solutions over others is known as a regulariser(Poggio *et al.*, 1985). Usually this model or *prior* is analytically defined, for example that the world is on average smooth (Horn & Schunck, 1981): it has a bounded and limited

differential[2]. Alternatively our model may be based on previous experience of the relationships between the measured variables in low noise cases. If we measured images in very favorable situations, and were certain that our measurements were correct, then based on a large amount of experience of this low noise data, we can form a model of the characteristics of accurate signals. If later, measurements are made in much worse conditions, then this knowledge of good solutions can be useful.

This model can then used in the interpretation of measurements which either have uncertainty, or where there are many possible interpretations of these measurements. This can be done by choosing out of the many possible potential solutions, the one that is most compatible with our model. Consider the case of distance estimates obtained from an analysis of stereo information. Often stereo only gives information as to the depth at a very few locations within an image, but sometimes, as when a surface is clearly textured, depth information is available through out an image of an object. If in all the cases where we have a information as to the depth throughout the image, we find that the depth varies smoothly, we can use this fact to infer that the depth will also vary smoothly in the cases where we only have limited depth information. By assuming smooth variation of depth between the sparse points, we can therefore infer probable depths for all the locations where there is in fact no data as to the depth: the model can be used to infer the state of the unmeasured world.

Consider another case where we have noisy pixel intensity estimates. This time, when the pixel intensity was measured in known low noise cases (usually when we have bright illumination), say we found that the local intensity varies only slowly within the image. With a noisy image, at least some noise can be removed by changing the estimated local intensity so as to force the local intensity to vary slowly as well. This means as well as being able to infer unseen measurements, a model can be used to remove noise.

Often the prior model is combined with a method so that no regularisation occurs where parts of the prior model are incompatible with a particular input (if the regulariser is that the world is smooth but the data says that there is a discontinuity at a particular location then don't smooth over that discontinuity (see (Terzopoulos, 1986),(Terzopoulos, 1988), (Geman & Geman, 1984)). Techniques based on regularisation have become widespread both in image reconstruction and in the computer vision

[2]This is a mathematical way of saying that we believe that the world is smooth. The differential measures how quickly something is changing, and if something changes in a discontinuity, then the differential will be infinite. By stating that we will only except interpretations where the differential is finite, we rule out all interpretations that include discontinuities and are therefore smooth

literature (Poggio *et al.*, 1985).

6.3 Regularisation as probabilistic inference

As well as acting as an auto–associative memory, the Hopfield architecture can be used to perform regularisation (Hinton & Sejnowski, 1983). To do this, we need both a method to represent a model of the world, and a means to combine this model with the input. One way we can view this is problem is probabilistically. If the world is represented in terms of the state of a number of binary features, then a model of the probability of every possible combination of feature states also constitutes a model of the world. If then as input, we get estimates of the probability of each individual features being present in the world, then we can combine the model and the input just using the laws of probability: the probability of any given combination of features is the probability of a combination of features given by the model multiplied by the probability of this combination based on the input.

Conceptually, this framework of assigning a probability to every combination of features based both on the its prior probability and the evidence for the features from the measured world allows both desired forms of regularisation. If we are certain of the state of a few of the variables, the model can be used to infer the most probable state of the other variables. If we have uncertain knowledge of the states of all the variables, the model can be used to estimate the most likely combination or state of these variables. Despite this, if we have a number of binary features N, the number of possible combinations of variables the we have to assign a prior probability to is 2^N. For even small visual problems, this is an unrealistic number to explicitly assign probabilities to, let alone to exhaustively search through. It is in this role of assigning probabilities to combinations of features, and searching for highly probable combinations that the Hopfield model can be of use.

6.4 Hopfield type networks for probabilistic regularisation

The process of probabilistic regularisation requires that a probability can be assigned to every combination features. This probability then has to be multiplied by the probability of this combination given an input (consisting of measurements of the world). Lastly we need a mechanism for finding probable combinations given both the model

and the input. The next section describes how this can all be done using the Hopfield network by using a different specification of the weights, and a different interpretation of its dynamics.

6.4.1 The Boltzmann machine learning algorithm: a method for learning a model of the world

The key to implementing probabilistic inference using the Hopfield network lies in the fact that when the update rule is not deterministic but probabilistic, then the network at equilibrium does not settle to a single state. Instead the probability of a given state (combination of features being true or false) conforms to a Boltzmann distribution where the probability of a given state is proportional to $\exp -E/T$ where E is the energy of the state (equation 6.3), and T is the temperature, a measure of the noise in the system (Hopfield, 1982). As can be seen from equation 6.3, the energy of the state is dependent on the weights and thresholds of the system. By manipulating these we can change the energies of all the feature combinations. By changing the weights in the network until the probability of each state of the network corresponds as closely as possible to the probability of the state being observed in the world, the network can be seen to be a model of the probability of each possible combination of features. How then do we find weights and thresholds that will make the probability of a state in the model correspond as closely as possible to that observed in the world? One method that does this is the Boltzmann machine learning algorithm (Ackley *et al.*, 1985).

A measure of the difference between the observed probability distribution of states P(S) in the world (where S is one configuration of output values for the feature detectors/neurons S_i, and the probability distribution of inherent in the model P'(S) is the Kulback Leiber distance:

$G = \sum_S P(S) \log (P(S)/P(S'))$

To find a good model, one method is to perform gradient descent in the measure G. To perform gradient descent in G is at least in theory simple. To do this we make measurements for the world in low noise situations and run the system twice. Firstly we set the outputs of the neurons in the network to their measured values and run the system to equilibrium using the Hopfield dynamics but with a probabilistic update rule (equation 6.2).. At equilibrium, we then collect statistics of the pair-wise correlations between the units: $P(S_i, S_j)$. This process is then repeated but this time without freezing the the activity of the units representing measurements from the world. Again we collect the pair-wise correlations $P'(S_i, S_j)$. If now the inter unit weights are

updated by an amount proportional to $P(S_i, S_j) - P'(S_i, S_j)$ then this will perform gradient descent in G. Barring local minima this will eventually result in a model of the world as close as possible to that observed. Therefore if we have access to a large number of measurements from the world that we trust, we can use this algorithm to set the weights in a Hopfield network so that the probability of any given combination of variables matches as closely to that observed in the world. The Boltzmann machine algorithm also works with "hidden" units, neurons involved in the dynamics but not directly associated with measurements from the world. Given enough hidden units, enough time, and enough samples from the world, it is potentially possible to model any probability distribution[3].

6.4.2 Using the Hopfield dynamics to combine a model with measurements from the world

Having a model of the world inherent in the network provides half the solution to performing regularisation. We also need to combine this model with measurements from the world, and to find the most probable solution. The desired behavior is that the probability of a state is just the probability of that state given our model of the world $P_{\text{prior}}(S)$ multiplied by the probability of it being generated given the measurements from the world $P_{\text{world}}(S)$ giving out final estimate of the states probability $P_{\text{posterior}}(S)$:

$$P_{\text{posterior}}(S) \propto P_{\text{prior}}(S).P_{\text{world}}(S) \tag{6.4}$$

$$\propto e^{-E_{\text{prior}}(S)}.P_{\text{world}}(S) \tag{6.5}$$

$$\propto e^{-E_{\text{prior}}(S)}.e^{-E_{\text{world}}(S)} \tag{6.6}$$

$$\propto e^{-E_{\text{prior}}(S)-E_{\text{world}}(S)} \tag{6.7}$$

This translates to adding a term to the energy of the network, representing the prior, so that the probabilities as determined by the new energy are equal to the posterior probability: that is the prior multiplied by the measured probability.

Given weights as set by a network such as the Boltzmann machine, estimates of the probability of the presence of features from the world (and hence an external input of η, and with the network initialised in a random state, the standard Hopfield dynamics are given a new interpretation. Since the dynamics always reduce (or leave constant)

[3]For awkward probability distributions with large numbers of variables, the time till the end of the universe may be a problem for such a direct approach.

the energy, the system will always move to more probable interpretations: the Hopfield model can act as a probabilistic regulariser.

In this way of viewing the Hopfield net, the previous inputs are viewed not as memories but as information to estimate the input probability distribution. The role is not to recall a single memory, but to use all of the previous experience to regularise the input. This provides a powerful framework for understanding recurrent nets.

6.5 Problems of the regularisation framework

Under the interpretation of the Hopfield net as a statistical regulariser, some of the old problems disappear. In the Hopfield network, there is the a severe problem of memory over load. If we attempt to store more patterns than the limited capacity of the network, then the network instantly ceases to operate as an auto-associative memory: the so called palimpset catastrophe. This problem can be seen as a virtue in the regularisation framework. The memories generated in the auto- association are generated from a random distribution. When enough samples have been seen so as to ascertain this, the correct behavior is to have no attractors (since because they are random, there are no more probable interpretations. Therefore the failure of the network to act as an auto-associative memory signals that it has ascertained the underlying statistical structure in the input: a virtue not a limitation. The "spurious" minima also cease to be a problem and can be seen as well motivated generalisations from the input statistics (MacKay, 1991b), and correlated non random inputs become useful rather than things to be avoided. Unfortunately these are replaced by three problems all relating to the relationship of the prior to the input.

6.5.1 Labeling the probability of an interpretation

In the "memory" formalism, all terminal attractors are assumed to be memories and therefore of equal status with no need to label the output as good or bad. In the regularisation framework, an output may be the most likely interpretation, but because the input is incompatible with the prior model, it may still be improbable, and the output should be contrasted with cases where the input is compatible with previous experience. The regularisation framework gives an explicit expression for the log posterior probability of a state S:

$$\log P_{\text{posterior}}(S) \propto -E_{\text{model}}(S) - E_{\text{world}}(S) \tag{6.8}$$

the log of the probability of an interpretation, given evidence from the world and the prior model of the world, is proportional to the network state's total energy. If the regularisation interpretation is used the probability (or estimate of it) needs to be know. This is especially true if the network's output is to be integrated with the outputs from more and less reliable networks.

Avoiding local minima

Secondly in the memorisation framework, it is the energy local minima that are of interest (these correspond to the memories). In the regularisation framework, where energy is directly related to the probability of interpretation, global minima are desired. Finding these is not generally possible in any realistic time and *the* global minimum for the given input is not required to make this system useful, but some technique to remove the worst of the local minima is required.

One minima avoidance heuristic is not to use the Hopfield dynamics at zero temperature, but use initially use a finite temperature (T) to search with. Because $P(S) \propto \exp(-E/T)$, when T is very large, all states are of near equivalent probability and there are no local minima. By slowly lowering the temperature, a process called simulated annealing, we can in theory avoid the worse local minima (given a lot of time (Geman & Geman, 1984)), and hence find solutions of high probability. This still leaves us the problem of choosing an appropriate temperature. This annealing can be speeded up if there are no hidden units in the system, by the so called mean field annealing. Here, instead of each unit firing on or off, each unit outputs a number between 0 and one based on its expected output ($S_i = 1/(1 + \exp(I))$ not 1 or 0). If there are no hidden units, the approximation is exact with the minima of a search based on the expected values being the same as that for the stochastic search. This can speed up the search by an order of magnitude.

One important difference from standard simulated annealing is that for regularisation, we wish to anneal only the model energy component based on the model. The energy component from the measurements has no local minima, and annealing this term would only add needless noise to it and hence slow the search. Instead it is better to only anneal the non-convex prior:

$$P_{\text{posterior}}(S) \propto e^{-(E_{\text{model}}(S)/T + E_{\text{world}}(S))} \tag{6.9}$$

This still leaves the two problems for these methods: some means is needed to select an appropriate temperature at any given time, and and some method is required to

communicate this temperature to the neurons.

6.5.2 Regularising only if the model is appropriate

Lastly, no matter how good the model is, there will be inputs that are new or different from previous experience. This is especially true is the model is simple. If we are confident of the data, but this data is incompatible with our model, then this indicates that the model is inappropriate for this data. Practical experience with regularisers has shown that when an input is presented that is very improbable under the model, but we are confident of the input, then that input should not be regularised since the result will tell you a lot about the model but little about the actual input. In the literature there are examples of methods to detect incompatibilities with the model and therefore prevent its application. Geman and Geman ((Geman & Geman, 1984)) use a simple model that states that nearby pixels are usually similar, but include a method for detecting areas in the image where this is not true (line processes) and use these to switch off the application of the model at these points. Terzopoulos (1986)) has a model that the depth of objects can be well approximated by smoothing splines, but again also has a method to detect where this model in inappropriate and therefore to not apply the model.

The probabilistic frame work can provide an alternative way of doing this. One measure of how appropriate the model is to the data is the probability of the currently estimated best state given just the model: $P_{\text{model}}(S)$. If this is low, then it is probable that the model is inappropriate for this data. Since T controls the relative contribution of the model and the data to the final probability, only if $P_{\text{model}}(S)$ is high should we anneal to the T=1. By controlling the final temperature based on a measure of the compatibility of the model to the data, we can avoid applying inappropriate models to the data. For this to be achieved, we need some measure accessible to all the neurons of $P_{\text{model}}(S)$ or equivalently $E_{\text{model}}(S)$ of the current state. For computer implementations this is easy, for a biological model it is not so obvious how this would be achieved.

6.6 Experiment 14

The preceding part of this chapter has argued that for regularisation, it is important to know the probability of a state or equivalently its energy. The aim of next experiment is to show that in a biologically plausible system constructed of spiking neurons, the energy of the system is directly represented in the temporal characteristics of the units'

output spikes. Specifically, if a network is in a low energy state and hence all firing neurons are connected via strong positive connections, the the neurons will tend to all fire at the same time: they will be phase locked. If the system is in a low energy state, with the neurons not connected strongly, the firing times will be essentially random. A measure of this phase locking (the amount of synchrony in the firing) can therefore be used as a measure of the systems energy.

6.6.1 Method

Architecture

In studying the effects of the energy on the degree of phase–locking, we have two effects that confuse the results. Firstly when being modified by the Hopfield dynamics, the energy of the network is constantly changing with each unit update. This effect can be removed by assuming units are either off or "clamped" strongly on. Since units that are off have no effect on the system[4], we need only simulate those units that are active. Here we model 100 such units which we assume are very probable given measurements from the world (they have large η, see equation ??), so that they are frozen on.

A second problem is that in cortex, the dynamics of the excitatory and inhibitory systems are very different having different time scales. To remove this complication, only the effect of the excitatory units was modeled (positive weights). This was justified partly because the time constants for the effects of the inhibitory system are longer, making them less significant to the phase responses of the cells [5]. Also the structure and connectivity of the cortex makes it likely that the second order statistics of the inputs are stored in terms of these excitatory connections, the inhibitory connections storing the first order statistics (a sort of local automatic gain control, see chapter 7, also (Amit *et al.*, 1987)). This is because the excitatory system both contains far more connections, and is much more specific in its connectivity. Since we are only interested in the model of second order statistics, only the excitatory system was modeled.

Since the input from the world (coming from the ηterms) is constant: the units were "clamped' external input to each modeled unit was constant and all units studied were on, the only variable contribution to the energy of the system state is the size of the lateral weights, those contributing to the prior energy component which equals:

$$E_{\text{model}} = -\tfrac{1}{2}\sum w_{ij} S_i S_j$$

[4] In a system with a 0-1 representation

[5] The $Gaba_A$ receptor, one of the receptors responsible for cell inhibition, is still effective after 100ms, not a good signal to coordinate locking on the scale of 5ms.

since in the net studied, all $S_i = 1$ (because of the the assumed evidence from the world, $E_{model} = -\frac{1}{2}\sum w_{ij}$. To set a net to have a particular model energy, all weights were initially chosen from a uniform distribution between 0 and 1, under the constraint that all $w_{xx} = 0$ and all $w_{xy} = w_{yx}$ [6]. To set E_{model} to some level C, the sum of all the lateral weights $\sum w_{ij}$ was found, and then each weight was divided by $2\sum w_i j/C$. This meant that the the model energy of the system was of a known value C.

In all experiments, 100 units were used. Two conditions were examined, full connectivity where every unit was connected to every other unit (but no self connectivity), and a more realistic diluted case where 90% of all connections were removed symmetrically (if the connection from unit 1 to unit 2 was removed, so was the connection from 2 to 1).

The units

The classical model of the action potential generation based on synaptic input is based on Hodgkin and Huxley's analysis of the operation of ionic channels in the neuron. Unfortunately to simulate such a system requires the simultaneous solution of a large number of non-linear differential equations, and simulating 100 such units of a length of time would be infeasible slow.. As a compromise between the realism of a full Hodgkin Huxley channel model of a neuron and the computational efficiency (with associated lack of realism) of the analogue Hopfield model (Hopfield, 1984), a two channel approximation proposed for biological reasons by Morris and Lecar was used (Rinzel & Ermentrout, 1989)). This model was used both because of its biological plausibility and because previous studies have shown it to have particularly good behavior for phase-locking (Cairns *et al.*, 1993). Both the author's unpublished studies, and previous work with similar models (Abbott, 1990) has shown that these models can implement characterised as performing descent in energy. The model was run under the parameter regime given in Rinzel ((Rinzel & Ermentrout, 1989) see appendix).

To aid interpretation, a time scale was associated with the net. An excited model neuron fires approximately every 150 time steps. If a highly excited neuron is assumed to fire at about 70hz, then this makes 1 ms equal to about 10 steps. This scale was only used to aid interpretation and has no functional significance.

[6] This was done because it is a requirement for assigning an energy functional to the Hopfield dynamics (Hopfield, 1982).

Units output

Biological cells in cortex mainly communicate through spikes. The Morris-Lecar model is a reasonable model of a biological neuron's membrane potential (with spiking behavior). We therefore communicate only spikes rather than the full membrane potential (see appendix).

Input integration

The effect of an excitatory input can be realistically modeled by an alpha[7] function with lambda equal to 2.5 ms ((Shephard, 1990) p 75). For computational efficiency, the finite rise time component of the alpha function was ignored, and was approximated as an exponential of time constant 2.5 ms (25 time steps) which is very fast to implement. On receiving a spike the input voltage is raised by an amount given by the connecting weight, it then decays back to zero in an exponential fashion.

For each unit there are two sources of input, the input η from the world, and the internal input via the lateral weights w. The external input is assumed to be a random process with a spike of weight 0.071 arriving with probability of 1/3 each time step. This is sufficient to clamp all units on with near to maximum activity (average input of voltage of 0.85), but with a reasonable variation in firing rate. This approximately corresponds to each neuron receiving input from 50 receptors, each firing in random phase at 70hz. The internal inputs come via the spikes of the other units moderated by the lateral weights.

Quantifying the spike arrival coherence: entropy as a measure of the spike arrival time probability distribution

To measure the coherence of the resulting spikes, the entropy of the spike emission times was used. This has the advantage over spectral methods in that it doesn't assume periodicity in the spike arrival, and over Gabor fitting to correlation maps as it avoids some of the biases introduced by such techniques. The entropy is an appropriate measure as it quantifies the uncertainty in the probability distribution: complete coherence (all spikes arriving at the same time) will give zero entropy. Complete randomness

[7]On receiving a spike, the membrane potential initially rises and this takes a finite time. After reaching its maximum, it then decays again. One function that also has this characteristic of initially rising and then falling is the alpha function: $f(x) = x^{\lambda}e^{-x/\lambda}$ where the first term dominates initially causing the function to rise, and then the second term comes to dominate causing the system to fall.

will give the maximal system entropy.

To calculate the spike arrival time entropy, the first 100 spike arrival times were recorded (each unit firing once), and these times were binned into 100 equal sized bins[8]. With the system continuing operation, the next spike from each neuron (a further 100 spikes) was recorded and the entropy of these calculated. This was repeated for the third group of hundred spikes (the third spike of each neuron). This allows the progress of coherence over time to be measured.

These estimates of the probability distribution will be confounded both by the effects of small sampling, and quantisation errors caused by being binned. To alleviate these problems, the smoothed binned distribution $\bar{B}$ was estimated from the binned estimates $B(x)$ using a symmetrical exponential smoothing kernel. The value $\frac{1}{2.5}$ was chosen experimentally so as to remove the random fluctuations in the completely random (unconnected) case. Periodic boundary conditions were used: time 0 was assumed to be close to time 99. This was done to remove phase effects dependent on when the main spike volley arrived (this is reasonable because the system is normalised to the average period of the system). The exact form of this smoothing is:

$$\bar{B}(x_i) = \frac{1}{100 \sum_{i=1}^{100} e^{-\frac{1}{2.5}|x-i|}} \sum_{i=1}^{100} e^{-\frac{1}{2.5}|x-i|} B(x_i) \tag{6.10}$$

where $|x - i|$ is the shortest distance between x and i assuming periodic boundary conditions, $B(x_i)$ is the number of spike measured in bin i, and $\bar{B}(x_i)$ is the smoothed binned estimate. The total sum of all the bins was found, and was found $P(x_i) = \bar{B}(x_i)/\sum_j \bar{B}(x_i)$.

The entropy of the smoothed binned probability distribution was then estimated as :

$S = -\sum_i P(x_i) log_e P(x_i)$

Since this expression is based on approximating the entropy of the continuous distribution via bins, there is an arbitrary constant, based on the size of the bins. To calculate the entropy of a completely phase-locked system, and a completely random one, these probability distributions were artificially created. The first was created by assuming all spikes arrive within the same time bin. The second was created by assuming that each bin contained only one spike. The gave an entropy of a completely random system of 4.61, and for a completely phase-locked system 2.60.

[8]The bin size was not specified in terms of time ticks because a measure of spike coherence that was independent of the spiking rate was desired By specifying the bin size in terms of the average firing rate, the measured coherence would therefore be the same independent of the firing rate.

Running the system

A trial consisted of selecting an energy = E. The weights were randomly generated and set to appropriate values as described previously. All units were run connected only to the external input for a random time between 100 and 600ms (equivalent to 100ms to 600ms time steps). This was done to randomise phase and remove start-up effects since the differential equations took some time to stabilise. The system was then connected and the entropy of the first, second, and third hundred spike arrival times was calculated. The system was run at energies between 0 and -80.0 at 0.25 intervals. For each energy level, we wish to ascertain how different weights and initial conditions effect the degree of phase coherence. To do this, for each energy level the system was run 64 times allowing both the average entropy and the standard deviation of the entropy to be calculated (a measure of the effect of different initial conditions and weights).

6.6.2 Results

The dependence of spike arrival time entropy on system energy was calculated for two systems, one with complete connectivity, and the other with 90% symmetrical dilution. The results are shown in figure 6.1.

To illustrate the phase-locking, two cases from the fully connected network were taken. One with a relatively high energy (figure 6.2), and one with a relatively low energy (figure 6.3).

The result of all these observations is that for a large range of energies, the degree of phase locking of the system is determined by the energy of the system. that the relationship is not that effected by either the initial conditions or the particular way the weights are configured (just there sum). This means that if a neural system has access to the degree of phase locking in the neurons, it also has access to a reliable measure of the systems energy.

6.7 Discussion

6.7.1 Output evaluation

In the regularisation framework described previously, the probability of a state given just the model $P_{\text{model}}(S) \propto -E_{\text{model}}$ and this probability is, as described before, an extremely useful quantity to know, especially if the output is to be combined with the

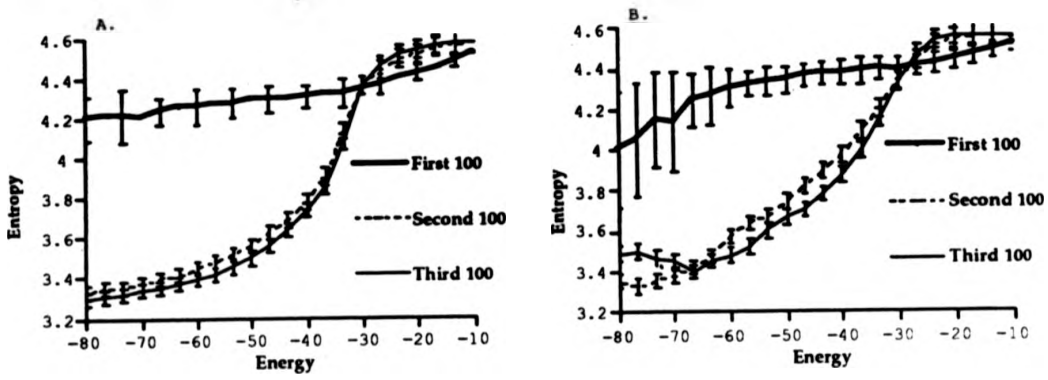

Figure 6.1: The entropy of the spike arrival time probability distribution plotted against the energy of the network. The entropies for the first, second and third spike arrival time are plotted (since there are 100 neurons, this is the entropy of the first 100 spikes). Figure A is for the completely connected network and B for the network with only 10% connectivity. An entropy of 4.61 represents completely random phase and an entropy of 2.60 represents complete coherence. The error bars correspond to 1 standard deviation of the distribution of entropies, and are only shown for every other data point for clarity. As can be seen from the diagram, as energy goes down (and under the regularisation framework, the probability of an interpretation goes up), the entropy also goes down (indicating that the system is phase locking). The small error bars indicate that the phase locking entropy is a reliable measure of energy, and is not highly dependent on the configuration of the weights: a system at energy -40 to be reliably discriminated from one at 35 within the second spike of the neurons. Note also the speed of phase locking: by the time the second spike arrives, most of the phase locking has already happened.

results of other nets. A module should always give its best interpretation (its most probable), but when the input is totally unlike any encountered before, due to noise, or due to coming from a previously stimulus completely unlike that encountered before, less weight should be given to this information. An interpretation that is improbable given previous experience is likely to be caused by noise or at least inferences made from it are likely to be suspect. By using phase coherence to communicate this, higher levels can weight the information appropriately.

A second interpretation of the energy is in terms of goodness of form: in Gestalt terms Prägnanz. The Gestalt laws of form : continuation, closure, proximity ..., can all be seen in terms of the input features obeying laws. If these laws are caused because

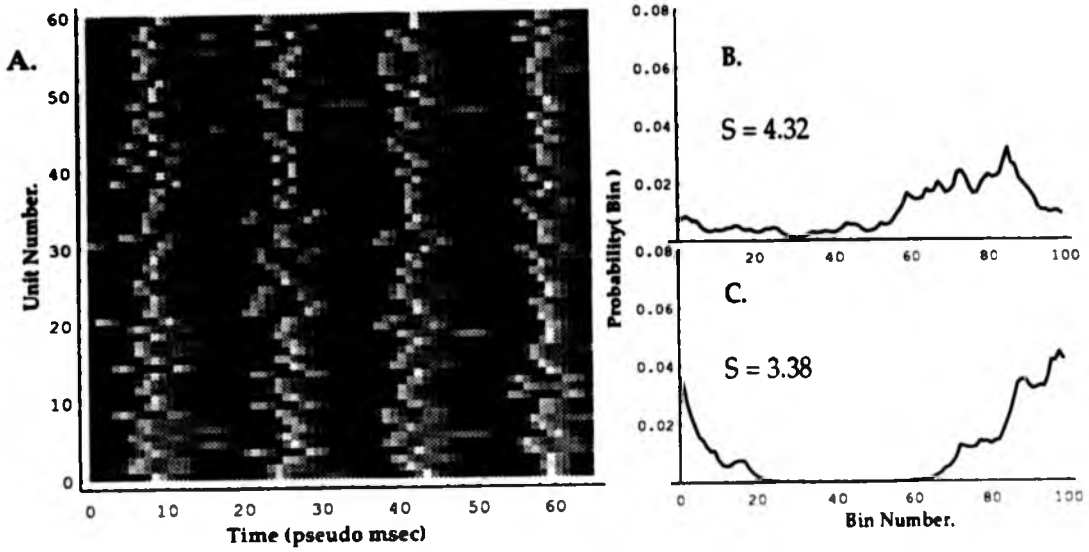

Figure 6.2: The evolution of spike arrival times with time in the fully connected network of moderate energy (-40.0). A) shows the evolution of the membrane potentials with time. B) shows the measured spike arrival time histogram for the first 100 spikes (from a different run). The entropy, S, measured from this distribution was 4.32. C) shows the probability histogram measured for the third 100 spikes, entropy 3.38.

these are common events in the world, then the energy of the system is a good measure of the coherence of the input, the Prägnanz. This experiment shows that the energy, and hence the probability of an interpretation is readily available in both a local form, in a form that is readily transmissible between networks (Eckhorn *et al.*, 1989), and in a form which neurons are sensitive to (Abeles, 1991).

6.7.2 Phase coherence for simulated annealing

The second problem with the regularisation framework is the need to avoid the worst local minima. The usual solution to the problem of finding global minima in nets is simulated annealing. Two problems with this scheme for a biological system are the need for an external temperature controlling homunculus to vary the units' noise (and hence effective temperature), and a scheme to choose an appropriate cooling schedule. Phase-locking provides a partial solution to these problems.

If each neuron has some intrinsic noise so that the net has a fixed T, then simulated annealing can be achieved by varying the effective contribution of the lateral

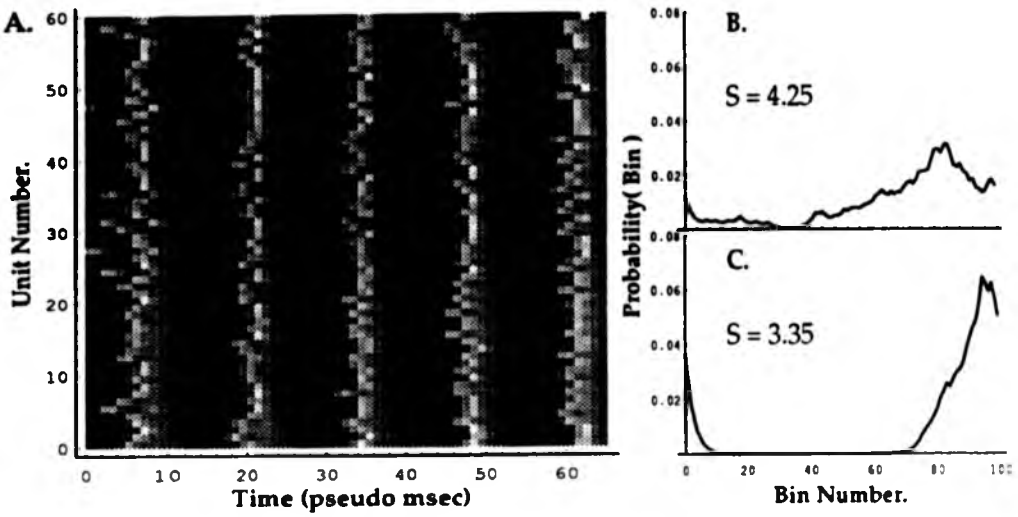

Figure 6.3: The evolution of membrane potential for the first 60 units in the fully connected network. The network was clamped to a state of low energy (-70.0). A shows the evolution of the membrane potentials with time. B shows the measured spike arrival time histogram for the first 100 spikes (from a different run. The entropy, S, measured from this distribution was 4.25. C) shows the probability histogram measured for the third 100 spikes (S=3.35). Notice that they are far more phase-locked than in the energy equals 40.0 case.

weights component (see equation 6.9). Phase–locking provides a means of doing this. Coherent activity is more effective in exciting a neuron than random activity ((Abeles, 1991) chapter 7). Therefore in the phase locked case the weights are effectively larger increasing their effective size. In the random phase case, the contribution of the weights is less and the dynamics are dominated by the noise, giving a higher effective temperature (see appendix). Therefore the degree of phase–locking can be used to control the effective size of the lateral weights and hence act as a temperature control.

The entropy of the phase–locking also provides a reasonable candidate for this temperature control. Setting the temperature to some quantity proportional to the (log) of the probability of interpretation seems intuitively sensible given a very limited annealing time. If the solution is already very good, adding noise will only make it worse. With a bad solution, adding noise could get you out of a local minimum. This is also done in a local manner, both in that the temperature is available locally at each unit, and in that there may be different levels of phase–locking (and hence different effective

temperature) at different parts of the net. Locally variable temperatures have proved useful in speeding convergence (Leinbach, 1989).

6.7.3 Appropriate application of the prior

The last problem is that if the input is very unlike the model of the world as expressed by the prior, then regularising it by the model will not improve matters. If our model assumes that the world is smooth and the input is of a very rough surface, then regularising it will just introduce distortions to the input: only if the input was likely to come from our model of the world should we use this model to clean up or regularise it.

Again by the method of doing simulated annealing by keeping the noise set and varying the effect of the weights, we achieve this. Since the effective strength of the prior is determined by the (log) of the probability of the interpretation (because unlikely models will not phase lock and hence will have a low effective contribution of the weights), then inputs unlikely to come from the model will not be regularised strongly by it. In standard simulated annealing, the output of the net at high temperatures is uninteresting. In this framework high effective temperature just means that the inputs have not been moderated by previous experience.

6.8 Conclusion

The cortex has large numbers of lateral connections and it is hypothesized that these are used to create a model of the world based on previous experience. Networks using lateral connections to store information about previous experience have traditionally been understood in terms of an auto–associative memory metaphor, but it is argued that regularisation provides a richer metaphor for the way knowledge of the model is applied to the input.

Given any model, one is interested in how compatible a particular input is with this model of previous experience. If an input is at variance with the model or there is just very little evidence for one interpretation over another, then this could indicate noise or just that the input is unlike anything encountered before (and therefore inferences are less likely to be true). Either way, a measure of the input's likelihood of occurring based on previous experience is useful.

The regularisation framework gives a way of quantifying this likelihood. The (log) likelihood turns out to be proportional to the energy of the system, unfortunately a

global quantity not directly accessible to the system. What has been shown is that in the system studied, the energy is directly reflected in the temporal coherence of the spike arrival times. Intuitively, this should be true of most phase locking systems (although not in such a simple form): a system with a large amount of excitatory connectivity will phase–lock quicker than one with little, and the amount of excitatory activity is directly related to the energy of the system.

This means that phase–locking can be viewed as a measure of the system's confidence in its output provided in a form that is locally accessible and capable of long range communication. It is also in a form that is useful for solving more technical problems relating to regularisation, the finding of good minima, and only regularizing inputs that are likely to have come from the model.

One of the functional roles usually assigned to the observed spiking coherence is that of feature linking (Singer, 1990). This places strong constraints on the activity. It requires identification of each unit with a particular phase and needs for this to be a reliable rather than statistical effect. Also for multiple labels, multiple independent phases are required and simulation studies show that getting more than three stable independent coherent groups is near impossible in the presence of noise.

The requirements for the probability communication hypothesis are much less strong. Most non linear oscillators will show stronger phase–locking when connected strongly together. By interpreting this not as a binary effect (phase locked or not) but as a variable effect, this opens up a very useful functional role for phase– locking, the communication of the confidence in a state, even if the requirements from the feature linking hypothesis are not met.

Chapter 7

A variable inhibition search algorithm for perceptual statistical inference

Summary

It has been proposed that inference based on previous experience is involved in low-level perception (Helmholtz, 1962; Nakayama & Shimojo, 1990; Nakayama & Shimojo, 1992; Barlow, 1990). If this inference is framed in the language of Bayesian statistical inference, as described in the previous chapter, this can be performed using simple binary neural networks (Hinton & Sejnowski, 1983; MacKay, 1991b). This chapter addresses three problems within this framework: the mapping of inference to the excitatory/inhibitory system found in cortex, the search technique used to find good solutions, and the communication of interim solutions. It is shown that if a modified covariance prescription is a used as an approximation for the weights. This form can be mapped simply onto an excitatory/inhibitory system with similar constraints to those found physiologically. The excitatory and inhibitory systems in cortex have two different time scales and this is used to suggest a method of search: variable inhibition search. This is tested on a number of sparse coded inference problems and shown to produce better (more probable interpretations) than Hopfield search (as proposed by MacKay), and is comparable or better than simulated annealing with a slow annealing time (as proposed by Hinton and Sejnowski). Whilst searching, interim estimates are communicated in a more usable form than by either the Hopfield or simulated annealing techniques.

7.1 Introduction

In a perceptual system, either man or machine based, that is trying to estimate the presence or absence of features in the world, using a model based on previous experience to perform inference is very useful. Even at the level of low-level features, edge detectors or disparity detectors, at least two particular forms of inference will be useful. When we have only sparse data, such as is often the case in estimating depth from stereo, then with a model of how depth varies with space based on previous experience with more complete data, we can infer the most probable depths of intermediate points (Nakayama & Shimojo, 1992). Model based inference is also useful in noisy situations, inferring the presence of edges in low lighting conditions for instance. Again, combining the noisy and incomplete estimates of the presence or absence of edges, together with a statistical model of lines (in the past they tend to be continuous, smoothly varying or whatever), allows the system to infer the most probable location of lines even when based on very limited evidence from the world. This process the Gestalt psychologists called completion.

The observation that the cells as early as V2 respond to the illusory contours in the Kanizsa triangle indicates that some inference is performed at a low-level (von der Heyt and Peterhans cited in (Zeki, 1992), see also Gilbert (1992b)). It has also been shown that in many illusions, low-level inferences predominate over high level inferences (Kanisza, 1979), and it has been suggested that low-level inference can explain stereo interpolation effects (Nakayama & Shimojo, 1992), and subjective contours (Nakayama & Shimojo, 1990).

In the previous chapter, it was shown that these problems of perceptual inference can be expressed in terms of probabilistic inference, and can be mapped on to a network of simple computational units. This chapter addresses the problems of mapping of the networks to more biologically plausible architectures, and finding a mechanism for performing the search for the best (most probable) solutions whilst conveying reasonable solutions throughout this search.

7.2 The prior or statistical model: approximations to the Boltzmann machine prescription

The previous chapter stated that one way to encode a probabilistic model into the weights was to run the Boltzmann machine learning algorithm. This has the advantage

of being able to learn not only a model based purely on visible units, but to infer hidden units that are useful for modeling the distribution of visible units. It is also in theory an exact formulation of the gradient (although in practice, the finite time spent sampling to find the inter statistics introduces quite large sampling errors). Unfortunately it is also an extremely slow method of learning. There are a number of methods to speed up this learning, by making simplifying assumptions about the input probability distributions. Although often the assumptions inherent in these calculations are incorrect, they at least usually provide an improved starting point for learning with the correct dynamics.

One set of simplifying assumptions is that the measured correlation between any two units is independent of the correlation as measured between any other units. This approximation, originally proposed by Hinton and Sejnowski (1983), allows the calculation of appropriate weights using only one pass through the data as long as the system has no hidden units. If the probability distribution is reasonably unstructured, this approximation can be quite good. An approximation of this form is also inherent in the mean field learning algorithm (Galland, 1993), and although this algorithm fails on some problems that the Boltzmann machine succeeds, it does works for many problems. Therefore, this approximation is worthy of study, if only as the weights found by such a system would make a reasonable starting point for later searching using the true Boltzmann machine learning algorithm[1] This approximation and the resulting weight prescription are also very similar to that later proposed by MacKay (1991) based on maximum entropy assumptions (the weight prescription is the same barring and additional additive term.

Given that an assumption based on the correlations between units being independent, to what value should we set the weights. Consider a network of binary feature detectors ($S_{i=1...N}$), each measuring from the world whether the feature is present ($S_i = 1$), or not ($S_i = 0$). Consider also we have a collection of examples of the world that we wish to use as the basis of our model. If we then calculate for each pair of units in the network i and j, the probability of each combination of pair-wise events $(P(S_i = 1, S_j = 1), (S_i = 1, S_j = 0), (S_i = 0, S_j = 1), (S_i = 0, S_j = 0))$, then by mak-

[1]This learning algorithm is a proposal for perceptual learning and in the learning of low-level psychophysical tasks as studied by Sagi, there appears to be two forms of learning. The first is very fast and usually most of the improvement occurs within 3-4 minutes. The second is slow, with it appearing that sleep and in particular REM sleep in between training sessions is required for this learning to occur. Previously Crick and Mitchison (1983), have proposed that REM sleep can be used to craft the structure of Hopfield memories and Hinton and Sejnowski (1986) proposed that this process could also be implicated in the operation of a Boltzmann machine. This initially rather implausible hypothesis at least seems more plausible in the light of Sagi's recent work on perceptual learning.

ing the independence assumption, the weights between the units should be (Hinton & Sejnowski, 1983):

$$w_{ij} = \frac{1}{2} \log \frac{P(S_i = 1, S_j = 1)P(S_i = 0, S_j = 0)}{P(S_i = 1, S_j = 0)P(S_i = 0, S_j = 1)} \tag{7.1}$$

and the threshold of the unit should be:

$$\theta_i = \frac{1}{2} \log \frac{P(S_i)}{(1 - P(S_i))} \tag{7.2}$$

This leads to a very simple way of performing inference. Numerous detectors are applied to the image and the probability of each feature being present (ignoring context) is estimated. This is used to apply an external input η to each unit (see previous chapter). The effects of prior knowledge are then applied by updating the units in the net in random order with this guaranteeing at least finding a local maxima of the probability of an interpretation. Alternatively the units are updated with noise, implementing simulated annealing in an attempt to find the solution of maximum posterior probability. This process will set to $S_i = 1$ on all features that are probable given both the input and previous experience and provides a simple, parallel and vaguely neural metaphor for low-level perceptual inference.

7.2.1 Problems for this as a biological system

Despite this, there are a number of problems of this system as a metaphor for the operation of the intra-area connections in the cortex:

1) The nature of the computational units. In the system all units have both excitatory and inhibitory connections. The cortex has a different structure: units are either excitatory with all connections increasing the activity of other units, or they are inhibitory with all connections inhibiting other units. This is at variance with the requirements that the weight between any two units having either a positive or negative effect dependent on the measured correlation between the two units.

2) Search. If the Hopfield dynamics are used for search (descent based search), then in the highly non–convex energy surfaces found in inference, the solutions will often be local minima. Mackay (MacKay, 1991b) p241 states of local minima that:

> Such states have been *inferred* to be probable states by the maximum entropy's generalisation from the measured statistics.

This is not true. The maximum entropy can infer as most probable states that have never been seen before, but, with the probability of a state being proportional to $e^{E_{total}}$, these unseen states should be global minima of the energy, the states with maximum probability. The local minima are dependent on the method of search, which the optimal solution (if unique) should not be. The maximum entropy connection proposed by MacKay is between the systems energy and the measured probability, and local minima of this energy may be more probable than near by states, but are not inferred to be probable (as we will see later, if the probability of a feature being true ($P(S) << 1$), then the local minima is often a solution with all units off. Given information from the world, this is not sensible).

The search method implicitly suggested in (Hinton & Sejnowski, 1983) is simulated annealing. This has the virtue of *eventually* avoiding local minima. Although this method has the virtue of generality, this form of search is notoriously slow. More importantly for much of the search time (when operating at high temperatures) the state signaled to higher levels will be essentially random. This is especially a problem when, as is likely in cortex, the prior probability of a feature being present is much less than 0.5 (sparse coding, only ≈10% of units are *on* (Abeles, 1991)).

Simulated annealing can be thought of as replacing the deterministic search on E by a search of the probability distribution measured by F, the free energy. The free energy is defined for the equilibrium probability distribution as F = E -HT (where E is the energy H is the entropy of the states probability distribution and T is the temperature). This means that for high temperatures, solutions of high entropy are favored, and this translates to solutions with an equal number of units on and off.

With sparse coding and finite temperature, the system will therefore spend a lot of its time in states with far too many units on[2].

3) Communication to other systems. If Hopfield dynamics are used, then the current most probable solution is what is communicated to other levels. Although reasonable this gives a very limited communication. Other states of near equal probability are never communicated to higher levels. In light of higher level context these may in fact be more plausible. The answer proposed in (Hinton & Sejnowski, 1983) of running the system at finite temperature also has limitations. The system will be continually changing its mind as to the most probable solution, and the other system will have to

[2]This can be understood with analogy to magnetism. Magnetization is caused by a bias in the orientation of the atoms spins. By heating the magnet, increasing the noise, the system is moved towards a state of high entropy with equal number of spins in both directions. It is consequently demagnetised.

observe the output over a while to find the networks most probable interpretation and even more to find the relative probability of different interpretations.

What is required is a system that while searching and not "sure" of the solution communicates all plausible predicates. As the search continues and better estimates of how probable each feature is present are found, the less probable predicates cease to be signaled. This would give the best of both worlds.

7.3 Mapping inference to excitatory and inhibitory units

7.3.1 The covariance approximation

The weight specification given in equation 7.1 is a non-linear function of the four conditional probabilities (although only three degrees of freedom). This will be complicated to implement as a synapse that expressed this weight could both be positive and negative, and need to store the 4 conditionals. This needs to be simplified.

To make further progress, it is proposed that we make two further assumptions about the probability distribution of the features. The first is that the average probability of a unit firing across the data set is (approximately) the same for all units ($P(S_i) \approx P(S_j)$ for all i and j). The second is that the correlation between any two units is neither very high or very low (under the Boltzmann prescription, the weight between any two perfectly correlated or anti-correlated units should be infinite; obviously a problem for a biological system). Give these two assumptions, a further approximation is proposed: that the weights (w'_{ij}) should be approximately equal to the covariance of the two units outputs:

$$w_{ij} \approx w'_{ij} = K.((P(S_i = 1, S_j = 1) - P(S_i = 1)P(S_j = 1))) \tag{7.3}$$

The quality of this approximation to the weight prescription given in equation 7.1 is shown for four different levels of bit probability ($P(S_i)$ = 0.5,0.4,0.3,0.2, for all i) in Figure 7.3.1.

7.3.2 Separating the two terms of covariance into the two biological systems

Given the covariance approximation, it is now possible to map the system to an excitatory and inhibitory system. Considering just the activation contribution by the lateral

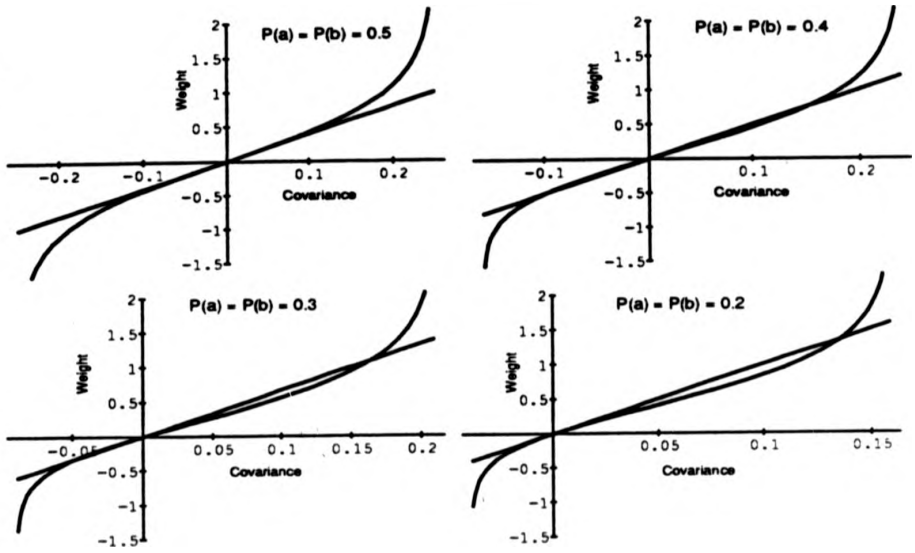

Figure 7.1: The match of the covariance approximation to the weight prescription given by by Hinton and Sejnowski (1983). P(a=1) and P(b=1) are the probability of each unit being on, and are fixed at four different levels (0.5,0.4,0.3, and 0.2), for the four diagrams. For each diagram, P(a,b) was varied and the resulting weight calculated both using the Hinton and Sejnowski prescription and the covariance approximation. The Hinton and Sejnowski prescription corresponds to the curved line, the covariance approximation corresponds to the straight line. In the Hinton and Sejnowski prescription, when the two units are always simultaneously on, the weight goes to infinity whilst that of the covariance approximation remain finite. Apart form this difference the two prescriptions are remarkably similar. Notice that for a given feature probability, the covariance can only vary between $-P(A)^2$ to $P(A) - P(A)^2$. In these diagrams, K was set to $\frac{4}{P(a=1)}$

weights (ignoring thresholds and external biases) this will be $I^{weights}$:

$$I_i^{weights} = \frac{K}{N} \sum_{j=1..N} (< S_i S_j > - < S_i >< S_j >) S_j \tag{7.4}$$

$$= \frac{K}{N} (\sum_{j=1..N} (< S_i S_j >) S_j) - \frac{K}{N} (< S_i > \sum_{j=1..N} < S_j > S_j) \tag{7.5}$$

Where K was empirically found to be just $\frac{4}{<S>}$ and N was the number of patterns that the statistics were collected over.. The first term $\sum_j (< S_i S_j >) S_j$ in a 1,0 representation is always non negative, this can be conveyed by the excitatory weights. The second term $(- < S_i > \sum_j < S_j > S_j)$ can be separated into two components. The sum of all unit activities weighted by their average activation is calculated. This will be

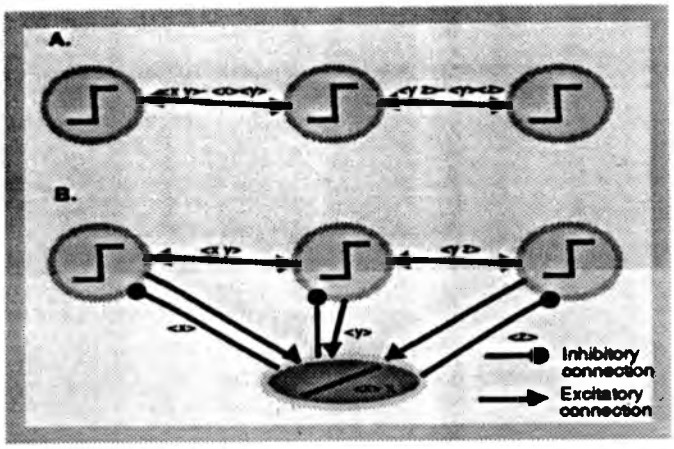

Figure 7.2: The two architectures. A The standard Hopfield network where all weights are proportional to the covariance (not all connections are shown). B The architecture implemented in terms of excitatory unit's together with a single inhibitory unit. The inhibitory unit sums all the units activities times the probability of them firing. This linear unit then inhibits all units again with a weight proportional to its probability of firing. Notice that in this scheme each unit will inhibit itself, an unwanted term. This disappears as the number of units $\rightarrow \infty$. Alternatively it can be removed by reducing the activity of the inhibitory unit by a small amount as done in the simulations.

the same for all units and can be calculated by one single linear inhibitory unit. This then inhibits every unit by an amount proportional to its activity as shown in diagram 7.3.2, it acts as a system wide automatic gain control.

7.3.3 Learning rules

Given this simple form for the weights, all can be learned with the same post–synaptic learning rule. Assuming that because of its high average activity, the inhibitory unit is always assumed to be on, this leads to the post–synaptic rule:

$$\delta w = O\alpha(I - w) \tag{7.6}$$

where O =1 if the post-synaptic unit is active (else 0), I equals 1 if the input unit is active (else 0) and α is the learning rate. This will make the weights converge to a smoothed estimate of the appropriate quantities. Notice that even if the weight change parameter is the same for all weights, the $< S_i S_j >$ will be on average $< S >$ times

	Real excitatory system.	Real inhibitory system.	Model excitatory system.	Model inhibitory system.
Relative number	80%	20 %	large	small
Response	non–linear	≈ linear	non–linear	linear
spontanous firing rate	low	high	low	high
δ w / w	??	??	large	small
Effective weight size	low	high	low $\propto P(on)^2$	high $\propto P(on)$
Comunication	fast (5ms)	medium (10ms) or slow (100ms)	??	??

Table 7.1: The different characteristics of the excitatory an inhibitory systems in the model and as experimentally found. δ w / w refers to the relative size of a weight change to the size of the weight.

smaller than those for both the inhibitory weights. This means that the ratio of the weight change to weight size will be much bigger for the excitatory-excitatory, than for the either the inhibitory-excitatory or excitory-inhibitory systems. This will make changes of the excitatory-excitatory much easier to detect. Whilst there have been many reports of synaptic modification in the excitatory-excitatory synapses, very few have been observed operating in the inhibitory system.

7.3.4 Similarities to observed physiology

Although the specifications of the weights and dynamics of the two model systems were specified only to be able to approximate statistical inference, the requirements are similar to the physiology as summarised in table 7.1. In the cortex, the ratio of excitatory to inhibitory units is about 4:1. For the model run with no noise, the ratio can be more extreme, theoretically only one inhibitory unit is required for the whole net. Given noise, and a limited linear dynamic range, more than one unit would be required. Despite this more excitatory feature units would be required than inhibitory level control units and a ratio of 4:1 is not unreasonable.

The model requires that the inhibitory units be linear, the excitatory units to be either threshold or sigmoid (for mean field annealing). Again the response of the inhibitory units as found in cortex is far more linear than the excitatory units. The spontaneous activity of the inhibitory neurons tends to be higher than that of the

excitatory units, again the inhibitory units in the model will be more active since they will fire when any of the excitatory units are firing.

The last aspect is the different time courses of the excitatory and inhibitory systems. In the model, the effect of the inhibition is negative. Of the two Gabageric inhibitory systems only one is subtractive, the $Gaba_b$ system. Unlike the $Gaba_a$ system which operates at medium to fast time scales (time to peak (ttp)$\approx$ 10ms) and is shunting or divisive, the subtractive inhibition caused by the $Gaba_b$ receptor operates on a much longer time scale than the excitation between cells (TTP $\approx$ 100ms as opposed to TTP $\approx$ 5ms). This contrasts strongly with the account of the role of inhibition presented by Amit and Treves (Amit & Treves, 1989) where the inhibition operates on a much faster time scale than the excitation. In our model the opposite occurs, initially the inhibition will be too low and many features that eventually will be turned off by inhibition will fire at the beginning of an inference. On initial exposure to a perceptual scene, a large number of units will fire many representing features not very probable to be present. Only after 100ms, when the $Gaba_b$ receptors have fully taken effect, will only the feature detectors representing features with a high probability be left.

The rest of this chapter explores this behavior, not as a handicap for the system, but as a possible solution to the problem of the search for a probable interpretation and the communication of this solution to later levels.

7.4 Experiment 15: variable inhibition for search

The next section describes a number of experiments to assess whether for systems with sparse representations, by reflecting the inferred biology and running the system with the initial inhibition too low, an effective search for inferences of high probability can be found. This form of search is compared to both Hopfield dynamics and simulated annealing.

7.4.1 Method

The architecture and model

The network consisted of 200 fully connected feature units, each having two states 0 and 1. The network implementing variable inhibition also had a single linear inhibition unit (see figure 7.3.2).

To perform statistical inference, a model based on previous experience is required. This was created by generating 50 random patterns. In each pattern the probability

of each feature being true ($P(S)$) was 0.25[3]. This created a probability distribution with some structure for the net to use to infer from. The thresholds for each feature unit were set using equation ??. For both the Hopfield and simulated annealing search techniques, the lateral weights were set to the covariance (equation 7.3). For the variable inhibition model, the excitatory connections were set to $< S_i S_j >$. A connection to and from the inhibitory unit was set to $< S_i >$.

To test the model for ability to infer the presence of features the following scheme was used. The system was told of the presence of twenty features. This was done by setting the estimated probability of that feature to 1 (from equation ?? this means that an external field of ∞ was applied: the unit was frozen on). The system was then asked to infer the most probable state of the other 180 features. Since the log posterior probability of a solution is proportional to -E (equation 6.3, the quality of a solution can be directly assessed by the solution states energy). The same set of weights and frozen units was used (an inquiry) with all three search techniques and the energy of the final solution assessed using the weights for the Hopfield model but ignoring external input.

7.4.2 The search methods

Three alternative search methods were used to find minima of the energy.

Hopfield descent method

The simplest technique: initially all unfrozen feature units were randomly set to be true with probability 0.5[4]. The search was carried out by asynchronously updating all units in a random order. Each unit above its threshold was turned on, all others turned off. This was repeated 50 times, each time in a different random order.

The Hopfield dynamics ensures that the energy of the state always either stays the same or gets lower ensuring that the final state is at least a local minima. This usually occurred within 10 iterations.

[3]This means that the coding system is sparse, for any given stimulus only a few features would be present. This has strong computational advantages in the amount of experience required to reliably specify the model (Gardener-Medwin & Barlow, 1992), it also appears that the cortex uses sparse coding (Abeles, 1991).

[4]They were set at P(S) = 0.5 to give the network a chance, if set with probability 0.25 as in the input patterns, then the network often settled to a state with all units off ((Buhmann & Schulten, 1989) shows this is an attractor for such a system). Usually this is a state of high energy.

Simulated annealing

The simulated annealing search is similar to that of the Hopfield model but with a non-deterministic update rule. This was performed in the following manner. The network was updated twice using the Hopfield dynamics. The average root mean squared activation of the units was calculated. This value was set to the initial temperature (T) ensuring that it was in a reasonable range. Each iteration was run as in the Hopfield model but the deterministic update rule was replaced with:

$$P(S_i = 1) = 1 - P(S_i = 0) = 1/(1 + exp(-I/T)) \tag{7.7}$$

where I is the summed input to the unit.

For each temperature the network was run for 3 iterations, then the temperature was divided by $\sqrt{2}$. This process was repeated 15 times resulting in 45 simulated annealing runs over a temperature range of two orders of magnitude. Finally the network was run for 6 iterations with T=0 (the Hopfield dynamics) to ensure that the final state was at least a local energy minima. This resulted in 53 iterations of the network, a reasonably long annealing schedule to compare to variable inhibition search.

Variable inhibition search

The variable inhibition search was meant to mirror the slower time scale of the inhibitory system compared to the excitatory system. This was achieved in the following manner:

- The net was initialised and run once with Hopfield dynamics to put it in a stable configuration.
- The activation of the inhibitory unit was calculated as the weighted sum of the activities. Its value was multiplied by a value to make it smaller. This value was initially 0.9. estimated activation.
- All units were updated in random order using the excitatory weights minus the modified inhibitory weight. After each update the value of the inhibitory weight was updated.
- After update cycle the value of the inhibitory unit multiplier was increased by 0.01. This was repeated 45 times till the final inhibitory multiplier was 0.945
- The system was finally run for 5 iterations with the inhibitory multiplier set to 0.945, this left the net at an approximately stable state.

This resulted in 50 optimization steps. Because of the splitting of the excitatory and inhibitory systems, detailed balance is not maintained and it is not guaranteed that this system will always perform descent, oscillatory behavior is possible. The starting inhibition multiplier was chosen such that the stable states had 50% of unfrozen units on. The final inhibition multiplier was not 1.0 because excitatory units inhibit themselves, an unwanted term. By setting the final inhibition to 0.945 it was empirically found that the effects of this interaction were avoided.

7.4.3 Results.

The time evolution of the search

The three networks were all given the same model and same set of frozen units and the temporal evolution of both the energy and the systems average activation was calculated for one problem. The results are shown in figure 7.3.

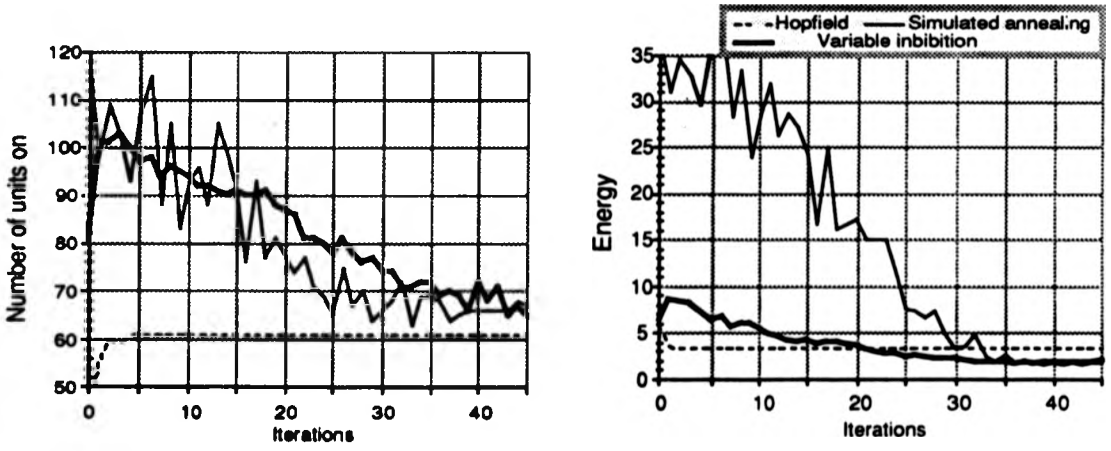

Figure 7.3: The evolution of the activation and energy ($\propto \log(P(interpretation))$) for one problem using the three different search techniques.

Three things can be seen from the graphs:

- With simulated annealing, for the majority of the time the quality of the solutions is very bad. This is because simulated annealing performs descent on free energy (E - TS where E is the energy, T is the temperature, and S is the entropy). All the good solutions in this sparsely coded network have low activity and hence low

entropy. Simulated annealing at a finite temperature, by emphasizing solutions of high entropy, moves the system away from good solutions.

- Although the number of units on with the variable inhibition search is large, the energy is comparable to that of the Hopfield through out the search.
- The final solutions found by both simulated annealing and variable inhibition search are better than that found by the plain Hopfield.

Results on a single problem

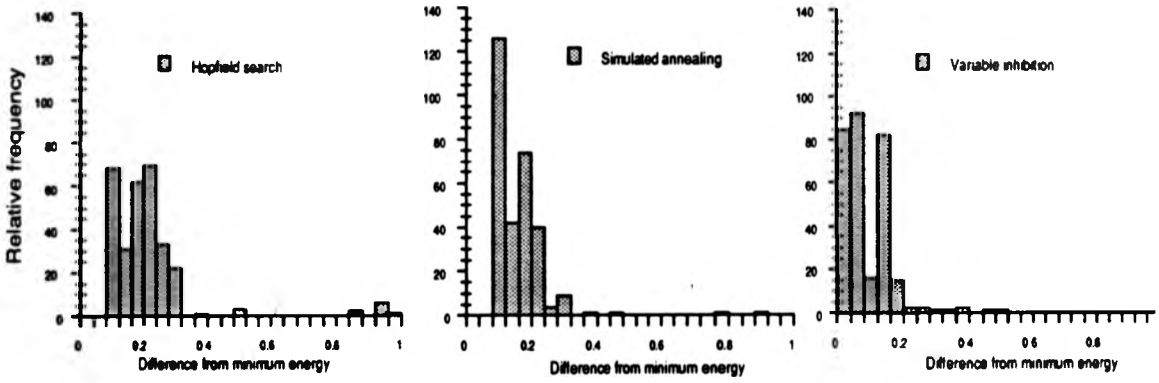

Figure 7.4: The probability density of the difference from the energy minima for one problem using the three different search techniques. A difference in energy corresponds to the ratio of the probability of each interpretation. Notice that the solutions for the Hopfield method are far more dispersed, this is because they are very dependent on initial conditions. Notice also that the only search technique that finds *the* minimum is the variable inhibition based search.

To get an idea of the relative characteristics of the three search techniques, one random inference was generated :a model together with a set of 20 true predicates, the model has to infer the most probable configuration of the other features. The three search techniques were applied to the problem 300 times. This gave us a reasonable estimate of the best solution (the minimum of 900 searches), the results of all three techniques could then be expressed as the difference from this minima. The results from two runs with different problems are shown in table 7.2. The probability distribution of estimates is shown in figure 7.4.

Things to be noted from the data are:

Condition	Average difference from minimum energy	SD difference from minimum	Minimum difference	Maximum difference
Annealing run 1	0.157	0.096	0.10	1.09
Hopfield run 1	0.214	0.159	0.10	1.04
inhibition run 1	0.082	0.069	0.00	0.49
Annealing run 2	0.217	0.096	0.077	0.66
Hopfield run 2	0.827	0.423	0.776	1.315
Inhibition run 2	0.222	0.100	0.0	0.54

Table 7.2: The results for two different problems averaged over 300 runs. For the first problem the best energy found was -1.24 by inhibition search; for the second the best solution found was 1.7056 again by inhibition search. This pattern of the Hopfield net performing better on low energy problems was repeated in other runs.

- The only method that found the most probable interpretation in both cases was the variable inhibition search.
- The average differences in energy are proportional to the geometric mean ratio of the probabilities of the solutions (from equation 6.3). This means that for problem 1, the solutions found by the variable inhibition method are 1.8 times more probable on average as those found by the Hopfield type search.
- The standard deviation of the Hopfield based estimates is much larger. It is not as important to find the best (most probable) interpretation as to always find a good one. The Hopfield method occasionally produces very bad solutions. This is because it is much more sensitive to initial conditions.

Results on multiple problems

The above results show that Hopfield based search produces results very dependent on initial conditions, and simulated annealing will communicate bad results for much of its search time. To see both if this is a result of the limited number of examples tried, and if there is a significant difference between the solutions found by simulated annealing and variable inhibition search, the three search techniques were tried on a larger number of problems in the following way: One hundred different models (collections of random patterns were generated. For each model, 25 different enquiries were made (randomly 20 units were frozen on) and each search technique was used to find the best solution. The quality of the solutions was again assessed via the final energy. The results are

	Simulated annealing	Hopfield	Variable inhibition
Average energy	1.52	1.87	1.46
Average activation	69.0	64.0	66.0

Table 7.3: The resulting energies and activations for 100 different sets of random patterns each test, with 25 different enquiries using the 3 different forms of search. Again the energy is $\propto$ the log of the probability of an interpretation.

shown in table 8.3.

There are two things to note from this data:

- The variable inhibition search produces lower energy states than the simulated annealing schedule used here.
- The final activation for the variable inhibition network is the same as that expected from the patterns (Expected activity = 20 (because of the unit clamping) + 180/4 (because in the patterns, 25% of units are on) = 65). This indicates that the model is working correctly, it is inferring states that are similar to those it has seen before.

7.5 Discussion.

7.5.1 Why variable inhibition works

The variable inhibition based search appears to be better than the Hopfield descent based search and at least comparable to the limited time simulated annealing schedule used here. One reason for better than Hopfield performance could be that for such sparse coded networks, when started with low activity levels, there is a large basin of attraction for the state of the network with all units off (Buhmann & Schulten, 1989). This was claimed to be useful behavior for an auto-associative memory system. For statistical inference this would be disastrous resulting in bad solutions for many inquiries dependent only on the random initial conditions of the unknown predicates. This behavior was found when the network was initialised with the bit probabilities the same as that of the input pattern, all unfrozen units often ended off. When the network was started with 0.5 bit probabilities this problem did not occur often as can be seen by the similar average final activities of the three networks in table 7.3.

The variable inhibition search can also be seen as similar to simulated annealing. By operating at an inappropriate activation, units that should be off are on introducing noise into the activations of all the units. By slowly decreasing the number of inappropriate units on, this noise term decreases. This is not the fast noise needed for simulated annealing, but the use of a variable level of signal to noise in a search technique is similar to simulated annealing.

7.5.2 Variable inhibition as a solution to the communication problem

As stated earlier, both Hopfield and simulated annealing systems have problems with communicating interim solutions whilst searching. The Hopfield model only communicates the currently estimated most probable solution and conveys no information about other plausible solutions during search. Simulated annealing not only communicates initially very improbable solutions (see diagram 7.3), but constantly changes its interpretation giving an initially very noisy estimates of potential solutions. In contrast, variable inhibition provides a reasonable solution to the problem of communication of interim results.

When search starts and the prior has not been applied, many features are potentially plausible. Since we don't know which will eventually be the most probable, and information from higher levels may make other interpretations more probable, we have to communicate all plausible predicates. As time continues, with the prior taking effect and higher levels having had access to other potential solutions, the network will have more confidence in which features are present or not. The threshold for being signaled as present can then be raised until only the features deemed to be most probable are signaled. This successive removal of plausible features as the networks estimate of the presence of absence of features gets better is what the variable inhibition search will do.

7.6 Conclusion

It has been shown that with sparse coding and moderate sized feature activity correlations, the weights required by a system to perform statistical inference are well modeled as the modified covariance. Adopting this simpler prescription allows the mapping of the inference architecture to an excitatory–inhibitory system. In the cortex these two systems operate at two different time scales. Rather than being a problem, this pro-

vides a simple solution to the problem of avoiding local minima in search that performs much better than the plain monotonic descent (Hopfield) and is comparable or better than finite time simulated annealing. By looking at the biology, a potential solution to the problem of search is found, the problem of different excitatory/inhibitory units not a problem but a solution.

Chapter 8

Summary and conclusions

This thesis has presented a number of ways in which representations can be generated, and models can be formed and searched. Each has been presented in isolation. In this concluding chapter all the previous findings will be reviewed. In conclusion it will be speculated on how these different problems can be integrated.

8.1 Principal components and the statistics of images

Chapters 2 and 3 on PCA and the related log contrast components started with the simple question: what are the principal components of natural images like? This question produced a number of findings, as follows,

1. The form of the early principal components.
2. An interesting match between the orientation tuning curves of two early "bar detectors" and measured human psychophysical performance.
3. The anisotropy in the components' tuning curves was shown to be caused by an anisotropy in the statistics of the world.
4. A related form of statistical analysis was proposed that avoided the problems of a large input dynamic range, and took into account that contrast is important, not the absolute measurements.
5. An analytical framework allowed an understanding of the form of the components.
6. A possible statistical reason was proposed for the increased orientation resolution as (opposed to scale) found psychophysically.

The first point is interesting because the components found were at variance with the form of the components predicted previously. It has been argued that all the components would resemble oriented spatial frequency filters (Daugman, 1990; Bossomaier & Snyder, 1986), and filters of this form had been "extracted" for a single image (Gonzalez & Wintz, 1977). The components found here did not all resemble those previously predicted. Whilst the early components do indeed resemble Gabors, the "cart wheel detectors" (filters that are radially symmetric as opposed to having reflection symmetry), are awkward for justifications of the Gabor preprocessing based on the Gabors supposed resemblance to PCA based operators.

The psychophysical match is interesting as it could indicate that low-level perception is tuned to the statistics of the images it processes. Other explanations could be found for the differences in horizontal relative to vertical sensitivity, the requirements of stereo, but this would argue for increased vertical resolution rather than horizontal as found.

The probable reason for this anisotropy is that the components 4 and 5 are the result of an admixture of three components, a vertical second derivative operator (horizontal bar detector), a horizontal second derivative operator (vertical bar detector), and a circular symmetric operator. Because of foreshortening there is more variance in the vertical direction and therefore this operator has a larger eigen-value. Therefore it is less mixed with the isotropic filter, giving it a more pronounced orientation tuning.

An anisotropy in the statistics of the world has been found because of foreshortening. The relative orientation tuning of the two "optimal" filters is determined by the anisotropy of the input statistics when tested on synthetic images. When the degree of anisotropy of the input statistics matches the anisotropy measured in the world, the anisotropy in the orientation sensitivities matches that found psychophysically by Foster and Ward (1991). That the filters in early vision are "tuned" by the statistics of the world provides a very simple explanation of this match. This is the main finding of Chapter 2.

The work on log contrast has one assumption built in, i.e., that there are global multiplications of the image (caused by illumination changes) which are not of interest. This simple observation results in a new preprocessing stage that provides a partial solution to the dynamic range problem. The resulting components are "simpler", and this technique could potentially be applied as pre-processing to other filtering schemes. The measured pre-processing in the retina holds out the possibility that a similar technique is also used in the brain.

8.2 The Statistical models of retinotopic space

8.2.1 The psychological representation of space

The simplest proposal is that if behavioral relevance was the driving force in the generation of topographic maps, then the perceived representation of space would correctly mirror that of the world. That we experience robust illusions shows that this is not true (the represented world has obvious anisotropies). Despite this the correlational theory is by no means unique in its prediction of such illusions as the horizontal-vertical illusion. Theories range from the proposal that it is an attempt to compensate for the foreshortening of the three dimensionality of the world, to the more fanciful explanation that at an early age, the father is both the most powerful thing in the child's world, and is (usually) viewed vertically (referenced in Underwood, 1966)).

What separates the correlational account from previous accounts of distance judgment distortions, is that it not only predicts an illusion, but also gives quantitative estimates of its size that match experiment. The observation that comparisons to $20° - 30°$ from vertical to horizontal give a larger distortion than a comparison of vertical to horizontal, is very awkward to explain in terms of a simple foreshortening argument[1]. The correlational account not only predicts this effect, but approximately the point of maximum illusion.

If the representation generated by the correlation was wildly inaccurate, then any creature using it would die. That small distortions cause little or no problems can be inferred from the way we cope. Accurately comparing lines of different lengths and orientations cannot be something that we do often otherwise we wouldn't have such steep learning curves for this task, and we would not initially be so bad at it as found in the performance of untrained observers (unpublished results by the author). By using assumed statistical regularities in the input, we lose the perfect mirroring of the structure of the world, but gain robustness and simple adaptability. Whilst flies, spiders, and snails, may have a hard-wired representation, for higher mammals, evolution may have opted for the virtues of robustness and adaptability rather than that of remaining faithful to the world.

8.2.2 The physiological model

The model with possibly the closest fit to empirical data found in this thesis was provided by the correlation based model of the geometry of V1. The complex logarithm

[1] Possibly the Freudian argument predicts this effect, if father was often drunk.

type representation has a number of good computational characteristics other than simply putting more representational emphasis on the point of fixation[2]. Therefore the logarithmic representation is not a unique prediction of the correlation theory. Despite this, the map cannot be purely logarithmic, as this would require an infinite representation of the fovea. This can be avoided (Schwartz, 1985) by adding a constant to the eccentricity before taking the log, but the author can see no way that in terms of invariance properties, this constant can be anything other than completely arbitrary.

This is not so with the correlation based theory, which produces a prediction of this value that is accurate to within 10%. Again the steepness of the logarithmic function is arbitrary from an invariance point of view, but is specified by the correlation proposal, and when matched against the power law exponent, the fit is within 10%. Lastly, the empirically found maps in cortex appears to be non conformal, the upper and lower fields are represented differently, with the vertical and horizontal meridians represented as different lengths. This non-conformal nature of the representation again causes problems for any purely invariance based theory. The correlation based theory predicts the degree of anisotropy of the lengths of the two meridians to within 5%.

The non-conformal nature of the representation should not be taken as an argument that the invariance properties of the approximate complex logarithm are not useful. Rather if the system is to be adaptable to the changing shape of the eye, using an explicit objective of creating a representation that possesses invariance properties presents problems. By implementing the system in terms of low-level input statistics, the problem becomes much easier. Fast and robust adaptation to the simple changes in space as occur for instance when you put your glasses on, becomes possible. For the sacrifice of an exact and conformal representation with perfect invariance properties (if these are useful), a robust, adaptable, and simple system is gained. It is proposed that this second choice is the natural one.

8.3 The computational uses of phase locking

Chapter 6 started with a simple observation that if things are strongly connected, they tend to behave similarly. Given this, it is not surprising given the strongly inter-connected nature of the cortex that cells tend to fire together. In vision this phenomenon tends to have been interpreted as a all-or-nothing phenomenon: the cells

[2]Both rotation and scaling are converted to translation, this has lead to the complex logarithmic representation being used in machine vision (Baloch & Waxman, 1991)

are either firing together or not. Here, the computational consequences of synchrony being a continuous phenomenon were investigated.

One framework where we would want to differentiate between strongly and weakly inter-connected states is that of Hopfield networks. Even if the system was working as an auto-associative memory, then because there often exist spurious memories (undesirable local minima) and these will be less strongly inter-connected than true memories, a continuous valued measure of the degree of connectedness of the final solution would give some warning that the state was not a true memory.

Despite this, as a framework for understanding low-level vision, auto-association seems an inappropriate metaphor and it is in the areas of cortex responsible for low-level vision that phase locking has mainly been reported. In this thesis, an alternative approach was proposed, that of forming and applying probabilistic models. Whilst it seems unlikely that the brain is in fact performing probabilistic inference, as a metaphor it is proposed as a better way of understanding its behavior. The proposal is also inherently testable. One such test would be to show two sets of stimuli: stereograms would seem an obvious choice. The first set would be consistent with being generated from an object smoothly varying in depth, the second set would be consistent with being generated by an object rapidly varying in depth. Since it is probable that most objects in the world are smooth, then the measured phase locking in cortex should be greater for the first stimulae than the second, and the degree of phase locking for a slightly less smooth object should be less than that for smooth stimuli but greater than that for the set varying rapidly in depth.

8.4 Variable inhibition search

The chapter on variable inhibition search was demonstrated on an abstract optimization task. Given enough time, simulated annealing will reach the best solution with probability 1 (Geman & Geman, 1984). The annealing schedule was also not highly optimized for the problem. The virtues of the proposed search technique were therefore not its absolute performance (although this was reasonable), but:

- Choosing an appropriate schedule is easy. For a sparse representation, if the inhibition is started so that 50% of the units are on, then decreased until the appropriate number of units are on, then this will be an appropriate annealing schedule. For simulated annealing, the points of phase transition should be found and this is very problem dependent.

- The output of the system is reasonable throughout the search process. This again contrasts with simulated annealing when for much of the time inappropriate solutions will be output.

- The search technique has a very simple implementation, only requiring the inhibitory units to have a slower time course than the excitatory units. This is consistent with known physiology.

The basic physiological requirement for the search is known to be present: the inhibitory system does take longer to operate than the excitatory system (Douglas & Martin, 1991). For evidence that this type of search is not just possible, but is used in cortex, we have to look at what would happen under a particular representation scheme. For low level perception, a representation in terms of oriented lines seems appropriate.

Andrews (1967a,1967b) investigated the perception of oriented lines with different presentation times. He proposed that his results could be understood in terms of a simple scheme using oriented filter units. Specifically, based on psychophysical experiments on the perception of small line segments, he proposed that:

> *Units vary in selectivity for orientation, being most selective when "tuned" near the horizontal and vertical*[3].

and that

> *Units integrate their outputs by a process of mutual inhibition which has a time-constant of the order* $\frac{1}{4}$-$\frac{1}{2}$ *second. Perceived orientation corresponds to maxima in the resulting pattern of inhibition*[4].

In the work on principal components, the resulting filters were horizontal and vertical. In the measured statistics of nature images, it was found that the correlation extended further in the horizontal and vertical axis than elsewhere. With these observations, if the visual system is tuning self to the statistics of images, then an increased sensitivity for horizontal and vertical is not unexpected: horizontal and vertical lines are more common than other orientations. What is of interest is what happens when this representation is used with variable inhibition.

What Andrews points out is that in such a system, when the inhibition is low and a large number of filters are on, the estimated orientation will suffer a systematic

[3] Andrews(1967a), p 995.
[4] Andrews(1967a), p 995.

bias. If the line element was at 10° then units with optimal tuning both clockwise and anti–clockwise will initially also respond, but because the orientation tuning curves are tighter for the filters near 0°, more units with orientation maxima > 10° will respond than those with orientation < 10°. This means that before the mutual inhibition has fully taken effect, the distribution of perceived orientation will be wide (a variation of perceived orientation of up to 30° was found for very short presentation times), and biased away from the horizontal and vertical axes (again this bias was found). This change in bias is precisely what would be expected if variable inhibition search is being used.

Again Andrews finds that the time course of this search is of the order of $\frac{1}{4}$–$\frac{1}{2}$ seconds. The time course for the effect of the $Gaba_b$ as measured *in vivo* in the cats primary visual cortex is 200–400ms (Berman *et al.*, 1991). Although Andrews gives this inhibition no functional role, he does make an observation that is compatible with its proposed role in this thesis. When integrating the outputs of a number of filters, the accuracy of combination is better than would be expected by purely logical combination, but can understood if:

> *The more accurate selection of response occurs only when the stimulus corresponds to a coincidence whose statistics have been stored.*[5]

To paraphrase, the use of statistical knowledge based on previous experience can be used to improve the interpretation of the outputs of the filter units. There seems no intrinsic reason that the inhibitory process should take so long, the brain has inhibitory receptor systems that operate much faster ($Gaba_a$ has a time course of $\approx$ 50ms). These experiments by Andrews are evidence that variable inhibition is operating. It also gives it a functional role. The combining of previous knowledge with an input using the technique proposed here is a non-convex optimization problem[6]. As shown in chapter 7, by using variable inhibition, good solutions can be found to such problems. The brain is possibly sacrificing the quality of the initial solution, so that prior knowledge can be applied, and the eventual solution found is better.

8.4.1 Adaptation in cortex: the violent policeman algorithm

Andrews also makes an observations about adaptation between the oriented filters:

[5] Andrews(1967b), p 1012.
[6] The problem contains local minima.

> The relation between presented and perceived orientation is subject to adaptation which tends to equalize the incidence of perceived contour orientation around the clock. The storage period supporting this adaptation is a matter of days[7].

For practical purposes, Gram–Schmidt orthogonalisation is usually used in neural network implementations of PCA. It is stable and converges better than alternative methods. The feature extraction component of PCA, (Hebbian learning together with some form of weight normalisation) is biologically attractive; a biological implementation of Gram–Schmidt orthogonalisation is less straight-forward. This led a number of researchers independently to propose an alternative means of orthogonalisation using inhibitory weights between the feature units (Rubner & Schulten, 1990; Rubner & Tavan, 1989; Földiák, 1989).

Although at first sight this scheme is more plausible, the number of inhibitory connections required scales badly with the number of feature units. This led to a proposal of the violent policeman algorithm[8]. This is a way of implementing PCA type representations without using Gram–Schmitt orthogonalisation, and without the need for very large numbers of inhibitory units and connections[9].

An attempt was made to remove the need for fully connected inhibitory weights. The network studied was one developed by Foldiak for deriving interesting information rich binary features. These only fired a small percentage of the time, thus producing a sparse representation (Foldiak, 1992; Toombs, 1992). The original demonstration used Hebbian learning of the weights, and an adaptive threshold. This was effectively applied to the multiple lines benchmark (Rumelhart & Zipser, 1985) but using biologically problematic fully connected inhibitory weights to prevent all units representing the same thing, or in other words, a form of orthogonalisation.

The requirement for fully connected inhibitory weights was removed by replacing them with a single inhibitory unit. This monitored the activity of the input units and extracted the most correlated group of units. This inhibitory unit, the "Policeman" unit, then pushed these inhibitory units apart.

The most correlated group of units was extracted using the Oja rule which, given zero mean inputs, extracts the first principal component. The "feature" units, being binary were not zero mean, but this process was sufficient to find the optimal solution

[7] Andrews(1967a), p 995.
[8] The credit for the interesting name goes to Graeme Mitchison.
[9] In the cortex, there are far more excitatory units and connections than inhibitory ones.

for the lines benchmark in about 75% of runs. This then provides a possible method for implementing the orthogonalisation required for feature extraction, and the behavior found by Andrews is consistent with such a mechanism operating.

8.5 Combining the three approaches: feature extraction and topographic map formation as model formation

This thesis deals with three different systems. One concerns the relationship between an image and a representation, one concerns the relationship between elements in a representation, and one concerns modeling in general. It is to be hoped that all three areas can be modelled using the same framework, and some progess has been made towards this.

Consider a neural network model where the hidden units are represented on a two dimensional grid, and each hidden unit is connected to all the input units, and to nearby hidden units. Such a network can be used to extract a topographic map if it is trained to find the weights of maximum posterior probability subject to the prior that units close together are highly correlated. The weights of maximum posterior probability can be found using a simple modification of the Boltzmann machine learning algorithm. The only modification required is in the calculation of the estimated correlation between the hidden units in the clamped phase: this is now calculated using a prior that nearby units are more highly correlated. Imposing this prior on the model means that if the inputs are locally correlated (as natural images are), then the representation found will be topographically organised.

As a model of topographic map formation, this framework has a number of strengths: 1) It relaxes the usual competitive assumption used previously in modeling these phenomenon, and allows not only winner-takes-all representations, but also distributed representations. 2) The framework is a general one, in which winner-takes-all, and distributed representation systems are special cases. It also makes clear what the network is doing: it is maximising the posterior probability of the weights. 3) By introducing controlled redundancy into the representation, it makes it more robust to internal noise. 4) The high local correlations make searching for good interpretations simple: the attractors are both deep and wide. 5) Although slow, the learning is simple, local, and approximately Hebbian.

A network of this form has been applied to 1 dimensional stimuli and has indeed

resulted in topographic maps (when the input statistics are locally correlated). When the network is also forced to find solutions where the probability of an individual neuron being on is small, then the receptive fields of indivual units resemble on-center off surround cells. Despite this, the network results in topographic maps where with a geometry that resembles that of the world (the mapping found is an identity mapping). As pointed out in chapter 5, the mapping in cortex is far from an identity mapping. The author is currently pursuing methods by which this problem can be solved. Hopefully a model that can potentially extract features, find topographic maps, and simultaneously find a probabilistic model of the world. There is still a lot of work to be done on this. The network has only been tested on one dimensional synthetic images, and the proposed modifications required to make it extract non-topographic maps require some crude segmentation of the images. Despite this, it is hoped that the probabilistic modelling approach can lead to a unifying of the three themes in this thesis.

8.6 Conclusion

In this thesis, it was attempted to both investigate the use of simple statistical mechanisms in vision, and to find evidence for these mechanisms operating in cortex. How successful was this enterprise? The first two chapters investigated the use of simple dimensionality reduction schemes for image description. Although it is not clear that a process exactly like PCA or LCPCA is used in low-level vision (in fact it seems unlikely), by using the statistics of natural images the results of Foster and Ward become very simple to explain. The two filters are aligned to the vertical and horizontal and vertical. Morover the horizontal bar detector has increased resolution over the vertical bar detector, and the ratio of the sensitivities is matched. All these aspects of the empirical data are easily explained with the simple premise that the operators are tuned to the signal they are to describe. That Andrews (1967a,1967b) reports that the statistics of testing appear to change the representation lends weight to this interpretation.

It could be that the operators are genetically predetermined, and the representation used by low-level vision fixed and unvarying. It could be that there are just by chance different sensitivities for the vertical and horizontal angle judgements. Such explanations are necessarily extremely *ad hoc*. It seems that any explanation that ignores the statistics of the world, but tries to account for such data, will be forced.

Previously physiological data has implied that correlation in the input is required to refine the representation of topographic maps. Previous models of topographic map

formation have assumed correlated input activity in order to learn maps in topographic correspondence with the input. Despite this, no previous work has looked at what the correlations in the world actually look like, and the implications for these correlation based models of the representation of space (with the exception of the thought experiments of Andrews (1964)).

In chapter 4, the correlation hypothesis is extended to psychological space with success. The horizontal-vertical illusion, its dependence on the environment, the maximal illusion not being for vertical but just off vertical, and the uncertainty in the representation being proportional to distance (Webers law): all these have been found experimentally in subjects, and all follow naturally from the correlation hypothesis. Other explanations have been given for individual effects, but none of the alternative explanations produce quantitative predictions, or are computationaly specifiable. Neither do they attempt to explain all the phenomenon, nor relate the physiological evidence to the explanation of psychological space.

It may not be Pearsons correlation that is used to recalibrate the low level representations. Craven (personal communication) has been using the number of zero crossings in the image when filtered with a Laplacian of a Gaussian filter as the measure of distance. This produces very similar predictions. Again, as with the first two chapters, it is not the exact method of using the statistics (correlation or zero crossings), but the idea of using image statistics at all. Doing so provides a very simple explanation for a number of phenomenon.

Chapter 5 turns the problem around. Previous network models have used correlated input statistics to create a representation that has the same geometry as the input. This is reasonable for modeling psychological space. For physiological maps such as in striate cortex, this is inappropriate as the geometry does not reflect that of the world. The main finding is that the correlations from the images studied match the structure of the striate cortex in macaque very well. This may be coincidence, but the match is very good. A comparison showed that the fall with distance is matched both in the form of the function but two parameters of the fall of this function. Interpreting this match is more difficult.

The proposal is that it is the point of fixation that is of interest, and points are represented only as far as they are relevant (correlated) to the point of fixation. This is possible but by no means completely satisfactory. As pointed out by Schwartz (1980), the complex logarithmic mapping has other advantages such as limited invariance properties. In animals that need to identify objects, more emphasis should be put on the

point of fixation. It could be that the exact form of the map is not important, but by using correlation, a usable signal is available to recalibrate the map. It may not be the exact form of the map that is important, but having a system that is robust. By calibrating using correlation with the point of fixation, the geometry can easily be recalibrated after damage, and robustness of the system could be more important to survival than any distortions induced by such a scheme.

The last two chapters discuss two aspects of using models to perform low-level inferences about the input. The first rests on a very simple property of interacting systems: if the components parts are strongly connected, the behavior of the parts will be strongly related (in this case the timing of the neurons spiking); if the components are weakly connected, the properties of the parts will be unrelated. The previous interpretation of coherence in the firing of cells is that it is a binary phenomenon: if the cells fire at the same time, the are representing features from the same "object". What is was proposed is that coherence can be more powerfully used as a graduated phenomenon, representing not the binary predicate *belonging to this object*, but the probability of the event given a model of low level features based on previous experience.

Whether this is actually occuring in cortex will require more experimental work, but it addresses an important problem in much neural network research. A neural network given an input, will always give an interpretation. Sometimes this will be good, sometimes bad. If a system has multiple sources of information, then it needs to know how reliable each interpretation is. For engineering based networks, there now exist methods for assigning approximate error bars to interpretations, but this is very computationally expensive (MacKay, 1991a). In a learning system, inferences made in regions where we have large amounts of data are going to be more reliable than those made from regions with little data. By signaling how likely the input was based on previous experience, and how often data has been received on inferences of this form, we will signal something related to the reliability of this interpretation. By signaling it in terms of the cells firing coherence, it is communicated in a form that can easily be interpreted by later cortical areas.

Chapter 7 addresses a problem if statistical models of this form are to be used in cortex. If a model is to be useful, it has to be applied quickly, and by the nature of these models, the search will be plagued by local minima. The solution proposed here was inspired by known physiology: the inhibitory $Gaba_b$ system takes a long time to settle (250–400ms), and is used in the visual cortex even although faster inhibitory neurotransmitters are present ($Gaba_a$). There is also limited evidence for variable

inhibition of the time course of 250–500ms from psychophysics (Andrews, 1967b). If the system was purely feed-forward and trying to communicate its solution as quickly as possible, then this can only be seen as bad engineering.

If instead the system is using a low-level statistical model to perform inference, then variable inhibition can be seen to have a role. This was demonstrated for a particularly difficult inference problem: given the model and told of the state of 20 of the features, infer the most likely state of the other 180. For this problem it was shown that indeed the variable inhibition search helped over simple gradient descent. It also didn't have the drawback of simulated annealing, where the initial estimates were essentially random. This process therefore gives a possible interpretation for the slowness of inhibition: by performing search in this way, we avoid local minima. Again whether search of this form is operating in cortex will require further experimentation, but the fact that the brain uses a slower time scale than necessary is suggestive.

This thesis started out with the hypothesis that because of the credit assignment problem, metaphors for computation based on simple statistical techniques would be important if low-level vision was to be learnt. It is hoped that the work which has been presented has indeed shown this. Even although the statistics used were simple, and each model had limitations, the results presented here are strongly suggestive of such processes operating.

Appendix

The Morris-Lecar neuronal model

The Morris-Lecar model is a two channel approximation to the dynamics of a single neuron where the membrane voltage v is dependent on the external input I in time dependent manner according to the following equations (Rinzel & Ermentrout, 1989):

$$\frac{dv}{dt} = -i_{ion}(v, w) + I \tag{8.1}$$

$$\frac{dw}{dt} = \phi \frac{[w_\infty(v) - w]}{\tau_w(v)} \tag{8.2}$$

$$i_{ion} = \bar{g}_{Ca} m_\infty(v)(v - 1) + \bar{g}_K w(v - v_K) + \bar{g}_L(v - v_L) \tag{8.3}$$

$$m_\infty(v) = 0.5 * [1 + tanh\{(v - v_1)/v_2\}] \tag{8.4}$$

$$w_\infty(v) = 0.5 * [1 + tanh\{(v - v_3)/v_4\}] \tag{8.5}$$

$$\tau_w(v) = 1/cosh\{(v - v_3)/(2 * v_4)\} \tag{8.6}$$

where

v	Voltage	$\bar{g}_{Ca} = 1.0$	$v_1 = -0.01$	$\phi = 0.33333$
w	Fraction of K^+ channels open	$\bar{g}_K = 2.0$	$v_2 = 0.15$	$v_K = -0.7$
I	Input	$\bar{g}_L = 0.5$	$v_3 = 0.1$	$v_L = -0.5$
			$v_4 = 0.145$	

The differential equations were integrated using a two step trapezoidal rule with a time step size δt of 0.09.

For the purpose of defining the spike train entropy and communication between cells, a definition of when a spike occurred was needed. This was defined as the first point where the differential of the membrane potential was negative after being positive, and that the potential was above 0.25 and had been below 0.0 sine the last spike. This complicated definition is required to avoid problems of the simulation equations ringing causing false spikes.

Simulated annealing via variably effective weights

The deterministic Hopfield model can be modified to perform simulated annealing by replacing the individual update rule by the following:

$$P(output = 1) = \frac{1}{1 - \exp(-I/T)} \tag{8.7}$$

where P(output = 1) is the probability of the output, I is the input, and T is the effective temperature (Hinton 1984). This can be achieved by just adding Gaussian noise to the input (because the logistic function is very similar to the cumulative normal) of the deterministic Hopfield network.

Traditionally the temperature of the system, the amount of noise added is varied, but since all that is of interest is the signal to noise ratio we can vary the signal keeping the noise constant. The signal I is:

$$I = \sum_i w_i^{effective} V_i \tag{8.8}$$

where $w_i^{effective}$ is the effective value of the i'th weight, and V_i is the output of the i'th unit.

If instead of varying the magnitude of the noise source, we vary $w_i^{effective}$, then we vary the effective contribution of the signal. If $w_i^{effective}$ is dependent not only on the weights size, but on its temporal relationship with other arriving spikes, then this can be used to vary the effective temperature of the system. Phase locking can be used as the temperature control for simulated annealing.

References

Abbott, L.F. 1990. A Network of Oscillators. *Journal of Physics A*, **23**, 3835–3859.

Abeles, M. 1991. *Corticonics: Neural Circuits of the Cerebral Cortex.* Cambridge University Press.

Ackley, D.H., Hinton, G.E., & Sejnowski, T.J. 1985. A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, **9**, 147–169.

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data.* , Chapman and Hall.

Amari, S. 1983. Field Theory of Self-Organising Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**, 741–748.

Amari, S.A. 1980. Topographic Organization of Nerve Fields. *Bulletin of Mathematical Biology*, **42**, 339–364.

Amit, D. 1989. *Modelling Brain Function.* Cambridge: Cambridge University Press.

Amit, D., Gutfreund, H., & Sompolinsky, H. 1987. Information Storage in Neural Networks with Low Levels of Activity. *Physical Review A*, **35**, 2293–2303.

Amit, D.J, & Treves, A. 1989. Associative Memory Neural Network with Low Temporal Spiking Rates. *Proceedings of the National Academy of Sciences, USA*, **86**, 7871–7875.

Andrews, D.P. 1964. Error Correcting Perceptual Mechanisms. *Quarterly Journal of Experimental Psychology*, **16**, 204–215.

Andrews, D.P. 1967a. Perception of Contour Orientation in the Central Fovea. Part 2: Spatial Integration. *Vision Research*, **7**, 999–1013.

Andrews, D.P. 1967b. Perception of Contour Orientation in the Centralc Fovea. Part 1: Short lines. *Vision Research*, **7**, 975–997.

Atick, J.J. 1992. Ecological Theory of Sensory Processing. *Network*, **3**, 213–251.

Atick, J.J., & Redlich, A.N. 1992. What Does the Retina Know About Natural Scenes. *Neural Computation*, **4**, 196–210.

Atick, J.J., Li, Z., & Redlich, A.N. 1990. *Color coding and its interaction with spatiotemporal processing in the retina.* Tech. rept. IASSNS-HEP-90/73. Institute for Advanced Study, Priceton.

Baddeley, R.J., & Hancock, P.J.B. 1991. A statistical analysis of natural images matches psychophysically derived orientation tuning curves. *Proceedings of the Royal Society B*, **246**, 219–223.

Baloch, A.A., & Waxman, A.M. 1991. Visual Learning, Adaptive Expectations, and Behavioral Conditioning of the Mobile Robot MAVIN. *Neural Networks*, **4**, 271–302.

Barlow, H. 1990. Conditions for Versatile Learning: Helmholtz's Unconscious Inference, and the Task of Perception. *Vision Research*, **30**, 1561–1571.

Barto, A.G., Sutton, R.S., & Anderson, C.W. 1983. Neuron Like Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**, 835–846.

Baylor, D.A., & Fuortes, M.G.F. 1970. Electrical Responses of Single Cones in the Retina of the Turtle. *Journal of Physiology*, **181**, 629–640.

Berman, N.J., Douglas, R.J., Martin, K.A.C, & Whitteridge, D. 1991. Mechansims of Inhibition in Cat Visual Cortex. *Journal of Physiology*, **440**, 697–722.

Blakemore, C., & Cooper, G.F. 1970. Development of the Brain Depends on the Visual Environment. *Nature*, **228**, 477–478.

Bossomaier, T., & Snyder, A.W. 1986. Why Spatial Frequency Processing in the Visual Cortex? *Vision Research*, **26**, 1307–1309.

Brown, L.G, & Shvaytser, H. 1990. Surface Orientation from Projective Foreshortening of Isotropic Texture Autocorrelation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 584–588.

Buhmann, J. Divko, R., & Schulten, K. 1989. Associative Memory with High Information Content. *Physical Review A*, **39**, 2689–2692.

Caelli, T., & Moraglia, G. 1986. On the Detection of Signals Embedded in Natural Scenes. *Perception and Psychophysics*, **39**, 87–95.

Cairns, D.E., Baddeley, R.J., & Smith, L.S. 1993. Constraints on Synchronizing Oscillator Networks. *Neural Computation*, **5**, 260–266.

Campbell, F.W., & Robson, J.G. 1966. Contrast Sensitivity, Fourier Analysis and Vision. *Journal of Physiology*, **186**, 551–566.

Carthy, J.D. 1958. *An Introduction to the Behaviour of Invertebrates.* Museum Street, London.: George Allen and Unwin Ltd.

Connor, F.R. 1982. *Noise.* Edward Arnold.

Cormack, E.O., & Cormack, R.H. 1974. Stimulus configuration and line orientation in the horizontal–vertical illusion. *Perception and Psychophysics*, **16**, 208–212.

Cowan, J.D., & Friedman, A.E. 1991. Studies of a model for the development and regeneration of eye–brain maps. *Pages 3–10 of:* Touretzky, D.S. (ed), *Advances in Neural Information Processing Systems*, vol. 3. Denver 1990: Morgan Kaufmann, San Mateo.

Craven, B.J, & Watt, R.J. 1989. The use of fractal image statistics in the estimation of lateral spatial extent. *Spatial Vision*, **4**, 223–239.

Crick, F., & Mitchison, G. 1983. The function of dream sleep. *Nature*, **304**. 111–114.

Daniel, M, & Whitteridge, D. 1961. The Representation of the Visual Field on the Cerebral Cortex in Monkeys. *Journal of Physiology*, **159**, 203–221.

Daugman, J. 1984. Two Dimensional Visual Channels in the Fourier Plane. *Vision Research*, **24**, 891–910.

Daugman, J. 1985. Uncertainty relations for resolution in space, frequency, and orientaion optimized by two-dimensional cortical filters. *Journal of the Optical Society of America,A*, **2**, 1160–1169.

Daugman, J. 1988. Complete Discrete 2-D Gabo Transforms by Neural Networks for Image Analysis and Compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **36**, 1169–1179.

Daugman, J. 1990. An information-theoretic view of analog representations in striate cortex. *Pages 403–423 of:* Schwartz, E.L. (ed), *Computational neuroscience.* Cambridge, Mass: MIT Press.

Dinse, H.R Kruger, K, & Best, J. 1990. A Temporal Structure of Cortical Informational Processing. *Concepts in Neuroscience*, **1**, 199–238.

Douglas, R.J., & Martin, K.A.C. 1991. A Functional Microcircuit for Cat Visual Cortex. *Journal of Physiology*, **440**, 735–769.

Duda, R.O., & Hart, P.E. 1973. *Pattern Classification and Scene Analysis.* New York: Wiley.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H.J. 1989. Coherent Oscillations - a Mechanism of Feature Linking in the Visual-Cortex - Multiple Electrode and Correlation Analyses in the Cat. *Biological Cybernetics*, **60**(2), 121–130.

Fawcett, J.W, & O'Leary, D.D.M. 1985. The role of electrical activity in the formation of topographic maps in the nervous system. *Trends in neuroscience*, **5**, 201–206.

Field, D.J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, **4**, 2379–2394.

Földiák, P. 1989. Adaptive Network for Optimal Linear Feature Extraction. *Pages 401–405 of: International Joint Conference on Neural Networks*, vol. 1. Washington 1989: IEEE, New York.

Foldiak, P. 1992. *Models of Sensory Coding.* Tech. rept. CUED/F-INFENG/TR 91. University of Cambridge,Department of Engineering.

Foster, D.H., & Ward, P.A. 1991. Asymmetries in Oriented-Line Detection Indicate Two Orthogonal Filters in Early Vision. *Proceedings of the Royal Society of London B*, **243**, 75–81.

Friedman, J.H. 1987. Exploratory Projection Pursuit. *American Statistical Association*, **82**, 249–266.

Gabor, D. 1946. Theory of Communication. *Journal of the Insitute of Electrical Engineers*, **93**, 429–457.

Galland, C.C. 1993. The Limitations of Deterministic Boltzmann Machine Learning. *Network*, **4**, 355–380.

Gardener-Medwin, A.R., & Barlow, H.B. 1992. The Effect of Sparseness in Distributed Representations on the Detectability of Associations Between Sensory Events. *Page 43 of: Proceedings of the Physiological Society*. University of Newcastle upon Tyne.

Gardner, E. 1987. Maximum Storage Capacity in Neural Networks. *Europhysics Letters*, **4**, 481–485.

Geman, S, & Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**.

Gibson, J.J. 1979. *The Ecological Approach to the Visual World*. Boston:Houghton Mifflin.

Gilbert, C.D. 1992. Horizontal Integration and Cortical Dynamics. *Neuron*, **9**, 1–13.

Gilbert, C.D., & Torsten, T.N. 1992. Receptive Field Dynamics in Adult Primary Visual Cortex. *Nature*, **356**, 150.

Gonzalez, R.C, & Wintz, P. 1977. *Digital Image Processing*. Addison–Wesley publishing company.

Goodhill, G. 1992. *Correlations, Competition, and Optimality: Modelling the Development of Topography and Ocular Dominance*. Tech. rept. CSRP 226. Sussex University.

Hancock, P.J.B., Baddeley, R.J., & Smith, L.S. 1992. Principal components of natural images. *Network*, **3**, 61–70.

Hebb, D.O. 1949. *The Organisation of Behaviour. A Neuropsychological Theory*. New York.

Helmholtz, Hermann von (translated by Southall, J.P.C.). 1962. *Treatise on physiological optics*. Dover.

Hertz, J., Krogh, A., & Palmer, R.G. 1990. *Introduction to the theory of neural computation*. Santa Fe Institute, Addison Wesley.

Hinton, G.E., & Sejnowski, T.J. 1983. Optimal Perceptual Inference. *Pages 448–453 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington 1983: IEEE, New York.

Hinton, G.E., McClelland, J.L., & Rumelhart, D.E. 1986. Distributed representations. *Pages 77–109 of:* Rumelhart, D.E., & McClelland, J.L. (eds), *Parallel Distributed Processing*, vol. 1, Foundations. Cambridge, Mass: MIT press.

Hopfield, J.J. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences, USA*, **79**, 2554–2558.

Hopfield, J.J. 1984. Neurons with Graded Responses Have Collective Computational Properties Like Those of Two-State Neurons. *Proceedings of the National Academy of Sciences, USA*, **81**, 3088–3092.

Horn, B.K.P, & Schunck, B.G. 1981. Determining optical flow. *Artifical Inteligence*, **17**, 185–203.

Hubel, D.H, & Weisel, T.N. 1974. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter and magnification factor. *Journal of comparitive neurology*, **158**, 295–306.

Hubel, D.H., & Wiesel, T.N. 1959. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology (London)*, **148**, 574–591.

Jones, J., & Palmer, L. 1987. An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, **58**, 1233–1258.

Kanisza, G. 1979. *Organisation in Vision : Essays on Gestalt Vision*. Praeger, New York.

Koenderink, J.J., & van Doorn, A.J. 1987. Representation of Local Geometry in the visual system. *Biological Cybernetics*, **55**, 366–375.

Kohonen, T. 1982. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, **43**.

Landau, P, & Schwartz, E. 1992. Computer simulation of cortical polymaps. *Neural Networks*, **5**, 187–206.

Leen, T.K. 1991. Dynamics of Learning in Linear feature discovery networks. *Network*, 85–106.

Leinbach, J. 1989. Automatic Local Annealing. *Pages 602–609 of:* Touretzky, D.S. (ed), *Advances in Neural Information Processing Systems*, vol. 1. Morgan Kaufmann Publishers Inc.

Levi, D, & Klein, S. 1992. Weber's Law for Position: the role of spatial frequency and contrast. *Vision Research*, **32(12)**, 2235–2250.

Linsker, R. 1986. From Basic Network Principles to Neural Architecture. *Proceedings of the National Academy of Sciences, USA*, **83**, 7508–7512, 8390–8394, 8779–8783.

Linsker, R. 1992. Local synaptic learning rules suffice to maximise mutual information in a linear network. *Neural Computation*, **4(5)**, 691–702.

MacKay, D.J.C. 1991a. *Bayesian methods for adaptive models.* Ph.D. thesis, California Institute of Technology.

MacKay, D.J.C. 1991b. Maximum entropy connections: neural networks. *Pages 237-244 of:* Grandy, W.T., & Schick, L.H. (eds), *Maximum Entropy and Bayesian Methods.* Kluwer Academic Publishers.

MacKay, D.J.C., & Miller, K.D. 1990a. Analysis of Linsker's Simulation of Hebbian Rules. *Neural Computation*, **2**, 173–187.

MacKay, D.J.C., & Miller, K.D. 1990b. Analysis of Linsker's simulations of Hebbian rules to linear networks. *Network*, **1**, 257–297.

Mallot, H.A. 1985. an overall description of retinatopic mapping in the cat visual cortex areas 17,18, and 19. *Biological Cybernetics*, **52**, 45–51.

Mallot, H.A. von Seelen, W., & Giannakopoulos, F. 1990. Neural mapping and space–variant image processing. *Neural Networks*, **3**, 245–263.

Marr, D. 1982. *Vision.* San Francisco: Freeman.

Meister, M, Wong, R.O.L, Baylor, D.A., & Shatz, C.J. 1991. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, **252**, 939–943.

Miller, K.D. 1990. Correlation based models of visual development. *Pages 267–353 of:* Gluck, A.M., & Rumelhart, D.E. (eds), *Neuroscience and connectionist theory.* Lawrence Erlbaum Associates.

Movshon, J. 1979. Two dimensional spatial frequency tuning of cat striate cortical neurons. *Society of Neuroscience Abstracts*, **9**, 799.

Nakayama, K, & Shimojo, S. 1990. Da Vinci Stereopsis: Depth and Subjective Occluding Contours from Unpaired Image Points. *Vision Research*, **30**, 1811–1825.

Nakayama, K, & Shimojo, S. 1992. Experiencing and Perceiving Visual Surfaces. *Science*, **257**, 1357–1363.

Normann, R.A., & Perelman, I. 1979. The effects of background illumination on the photoresponses of red and green cones. *Journal of Physiology*, **286**, 491–507.

Obermayer, K. Ritter, H., & Schulten, K. 1990. A principle for the formation of the spatial structure of cortical feature maps. *Proc.Nat.Acad,Sci,U.S.A.*, **87**, 8345–8349.

Oja, E. 1982. A Simplified Neuron Model As a Principal Component Analyzer. *Journal of Mathematical Biology*, **15**, 267–273.

Pentland, A.P. 1993. Surface Interpolation Networks. *Neural Computation.*, **5**(3(5)), 430–442.

Pettet, M.N, & Gilbert, C.D. 1992. Dynamic changes in receptive field size in cat primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, **89(17)**, 8366–8370.

Plumbley, M.D. 1991. *On Information Theory and Unsupervised Neural Networks.* Tech. rept. CUED/F-INFENG/TR.78. Cambridge University Engineering Department.

Poggio, T., Torre, V., & Koch, C. 1985. Computational Vision and Regularization Theory. *Nature*, **317**, 314–319.

Pollock, W.T., & Chapais, A. 1952. The apparent length of a line as a function of its inclination. *Quarterly Journal of Experimental Psychology*, **4**, 170–178.

Prestige, M.C., & Willshaw, D.J. 1975. On a role for competition the formation of patterned neural connections. *Proceedings of the Royal Society of London B*, **190**, 77–98.

Rinzel, J., & Ermentrout, G.B. 1989. *Analysis of Neural Excitability and Oscillations : In Methods in Neuronal Modelling - From Synapses To Networks Ed. Koch,C and Segev,I.* M.I.T. Press.

Ross, H.E. 1990. Evironmental influences on geometrical illusions. *Pages 216–221 of:* Muller, F (ed), *Frechner Day 90: Proceeding of the 6th annual meeting of the international society of psychophysicists.*

Rubner, J., & Schulten, K. 1990. Development of Feature Detectors by Self-Organization. *Biological Cybernetics*, **62**, 193–199.

Rubner, J., & Tavan, P. 1989. A Self-Organizing Network for Principal-Component Analysis. *Europhysics Letters*, **10**, 693–698.

Rumelhart, D.E., & Zipser, D. 1985. Feature Discovery by Competitive Learning. *Cognitive Science*, **9**, 75–112. Reprinted in (Rumelhart *et al.*, 1986, chapter 5).

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Vol. 1. Cambridge: MIT Press.

Samuel, A.L. 1967. Some studies in machine learning using the game of checkers II-recent progress. *IBM journal. R and D*, **11(6)**, 601–617.

Sanger, T.D. 1989a. Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network. *Neural Networks*, **2**, 459–473.

Sanger, T.D. 1989b. An Optimality Principle for Unsupervised Learning. *Pages 11–19 of:* Touretzky, D.S. (ed), *Advances in Neural Information Processing Systems*, vol. 1. Denver 1988: Morgan Kaufmann, San Mateo.

Schwartz, E. 1985. On the mathematical structure of the visuotopic mapping of macaque striate cortex. *Science*, **227**, 1065–1066.

Schwartz, E.L. 1980. Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research*, **20**, 645–669.

Shatz, C.J. 1992. The developing brain. *Scientific American*, **267**, 34–41.

Shatz, C.J., & Stryker, M.P. 1978. Prenatal tetrodotoxin infusion blocks segregation of retinogeniculate afferents. *Science*, **242**, 87–89.

Shephard, G.M. 1990. *The Synaptic Organisation of the Brain.* Oxford University Press.

Singer, W. 1990. Search for coherence: a basic principle of cortical self organisation. *Concepts in Neuroscience*, **1**, 1–26.

Srinivasan, M.V., Laughlin, S.B., & Dubs, A. 1982. Predictive coding: A fresh view of inhibition in the retina. *Proc.R.Soc.London Ser.B*, **216**, 427–459.

Sutton, R.S., & Barto, A.G. 1991. Time Derivative Models of Pavlovian Reinforcement. *In:* Gabriel, M., & Moore, J.W. (eds), *Learning and Computational Neuroscience.* Cambridge: MIT Press.

Terzopoulos, D. 1986. Regularization of Inverse Visual Problems Involving Discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(4), 413–424.

Terzopoulos, D. 1988. The Computation of Visible-Surface Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4), 417–438.

Toombs, P. 1992. *Representation of multi-scale images of words.* M.Phil. thesis, University of Stirling, Centre for cognitive and computational neuroscience., Stirling, Scotland.

Tootell, R.B, Silverman, M., Swikes, E., & Valois, R. De. 1982. Deoxyglucose analysis of retinatopic organisation in primate striate cortex. *Science*, **218**, 902–904.

Tootell, R.B., Silverman, M.S, Swikes, E., & Valois, R. De. 1985. Reply to Schwartz. *Science*, **227**, 1066.

Underwood, J.B. 1966. *Experimental Psychology.* Meredith Publishing Company.

Van Essen, D.C. Newsome, W.T., & Maunsell, J.H.R. 1984. The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Research*, **24**, 429–448.

Vidyasagar, T.R., & Henry, G.H. 1990. Relationship between preferred orientation and ordinal position in neurones of cat striate cortex. *Visual Neuroscience*, **5**, 565–569.

von der Heydt, R. Peterhans, E., & Baumgartner, G. 1984. Illusory contours and cortical neuron responses. *Science*, **224(4654)**, 1260–1262.

von der Malsburg, C, & Willshaw, D. 1977. How to label nerve cells so that they can interconnect in an ordered fashion. *Proceedings of the National Academy of Sciences, USA*, **74**, 5176–5178.

Watt, R.J. 1987. Scanning from coarse to fine scales in the human vision system after the onset of a stimulus. *Journal of the Optical Society of America. 4A*. 2006–2021.

Watt, R.J. 1991. *Understanding vision.* Academic Press: London.

Whitelaw, V.A., & Cowan, J.D. 1981. Specificity and plasticity of retinotectal connections: a computational model. *Journal of Neuroscience*, **1**, 1369–1387.

Willshaw, D., & von der Malsburg, C. 1976. How patterned neural connections can be set up by self organisation. *Proc.R.Soc.London Ser.B*, **194**, 431–445.

Willshaw, D., & von der Malsburg, C. 1979. A marker induction mechanism for the establishment of ordered neural mappings: its application to the retinotectal problem. *Proc.R.Soc.London Ser.B*, **287**, 203–243.

Yuille, A.L. Kolodny, J.A., & Lee, C.W. 1991. Dimension Reduction, generalized deformable models and the development of ocularity and orientation. *Pages 597-602 of: International Joint Conference on Neural Networks*, vol. 2.

Zeki, S. 1992. The visual image in mind and brain. *Scientific American.*, **9**. 43–50.

Zeki, S.M. 1978. The cortical projections of foveal striate cortex in the rhesus monkey. *Journal of Physiology*, **227-244**, 277.

Zhang, J. 1991. Dynamics and formation of self-organising maps. *Neural Computation*, **3**, 54–66.