Eccentricity Bias Models of Object Representation

by

Erick Chastain

Senior Thesis

Presented to the Faculty of the Undergraduate School of

Carnegie Mellon University

in Partial Fulfillment

of the Requirements

for the receipt of

School of Computer Science Honors

Carnegie Mellon University

August 2006

To professor Plaut, my mother, and Mark Stehlik for their boundless wisdom, support, and guidance.

Acknowledgments

Thank you to Professor Plaut for his infinite patience and helpful guidance. Lens was being an evil beast because of how gcc handles floating-point numbers, and he helped me to find this. Also, for all of the articles and discussions about everything from learning algorithms to the nature of cognitive neuroscience. Finally, his unerring support for me in ventures that were not strictly within the purview of the thesis work, notably anything ”Bayesian” or involving statistics. Secondly, to my family for listening to my theory-of-thmoment for the last 6 years and helping me keep on-track. Finally, for Mark Stehlik, who still has more books than I do... and is the chief reason why this document or its author even exist as they do today.

ERICK CHASTAIN

Carnegie Mellon University

August 2006

Eccentricity Bias Models of Object Representation

Erick Chastain, SCS Honors

Carnegie Mellon University, 2006

Supervisor: David Plaut

...

Contents

Chapter 1

Eccentricity Bias and Human Object Representation

1.1 Introduction and Motivation

Sometimes we take for granted our ability to recognize faces. It is one of the basic needs for social interaction, and it is effortless. How well could we relate to another person if we never recognized their face? In fact, we can recognize faces under a dizzying number of challenging circumstances. Intuitively, unconsciously, we can recognize faces, even on a foggy San Francisco morning.

As easy as it seems for humans, it is actually quite difficult. After all, faces look quite similar from a distance. As with many problems that the visual system solves efficiently, an understanding of face and object recognition has eluded the computer vision community even now. Yet in the early days of artificial intelligence, the problems of the visual system were seen to be easy, because they are problems that we are so good at solving. Famously, Marvin Minsky once assigned one of his graduate students the task of solving vision in a summer. The summer has been long, and robust object recognition has eluded the computer vision community.

Similarly, in the neuroscience community face and object recognition are hotly-contested and filled with many divergent viewpoints. For example, one theory is that there are face-selective areas of the brain, which can be selectively impaired by brain damage. This is an interpretation of activation and double-dissociation studies . Thus one of the voices in the choir is that face processing is handled by a specialized module, with faces being fundamentally different than other objects. However, other evidence suggests that the so-called face areas are actually used to recognize individual identities of objects within a category with which the subject has expertise.

There is however a theory that could explain all of this evidence. Faces are grossly similar, and as such face recognition may require fixation and high-resolution information about

subtle differences in hairline, nose structure, and eye-color. In other words, visual acuity may be needed to recognize faces. In the brain (occipital lobe), central and peripheral visual regions are spatially distributed. Interestingly, face images activate center-biased regions, whereas buildings activate periphery-biased regions. Of course, the peak of activity shows this stereotyped pattern. The activation recorded in fMRI studies actually has activity distributed in a way that falls off from the peak as a function of distance.

In the traditions of psychophysics and neuropsychology, an ideal model of face and house recognition is built to test this theory and evaluated by comparing its performance to findings from empirical studies on subjects with brain damage. Using a topographic bias which favors short, local connections and devoting more resources towards faces than houses, the model develops two overlapping areas to represent faces and houses. The face area is biased towards the center, and the house area is biased towards the periphery. Selective lesions to the center impair face recognition while leaving house recognition unaffected. Selective lesions to the periphery impair house recognition while leaving face recognition unaffected. A small number of unique neurons, which suggests that the visual system has a representation of faces that is optimal in coding length. House recognition is actually shown to require a significantly more holistic representation, requiring largely the same cells, but differing in activation pattern. Finally, methods to analyze causal relationships and new methods to create graphical models from energy functionals are exploited to understand something quite fundamental about the nature of face recognition.

In the analysis of such a complicated task, and what makes the human brain so adept at it, many new methods are developed, as a culmination and combination of other methods. New ways of analyzing optimization problems are at the core of this analysis, as well as ways to analyze complex, nonlinear, and noisy systems. Finally, we develop a method to optimize arbitrary multivariate energy or loss functions in n variables by converting them into

graphical models and using loopy belief propagation. As graphical models, energy functionals whose optimization requires belief propagation on graphs with tight cycles are more difficult to optimize. However, it can be done, if ultimately with caution. This approach yields a graphical model which shares the same assumptions and learning behavior as our model, but is more tractable computationally and more general. Furthermore, by making a graphical model of the task, the task is made algorithmic and analytical, and constitutes a suggested algorithm for face and object recognition which is consistent with neural representations of faces and houses. Future work is suggested to implement the algorithm and test its performance on the FERET database.

1.2 The Neural Representation of Faces and Objects

Face recognition is important but, as commented, highly controversial. In a sense, it is the premier battlefield for old battles that loom large over the field of cognitive science and philosophy of mind. The root of this controversy is the organization of the ventral, temporal visual system. Some claim that objects are represented with specialized modules for every category (Kanwisher, 2000), each with different function. This is a very old idea, dating back to Aristotle, but described most eloquently by Jerry Fodor (Fodor, 1983). Others claim that objects are represented with a distributed, overlapping neural instantiation (Haxby and others, 2001); (Ishai et al., 1999); (Pietrini et al., 2004). The controversy then is whether the ventral temporal cortex is domain-specific. That is, whether we are using what amounts to the fanciest cookie-cutter in the world to represent objects.

An additional perspective is that identifying specific individuals in an object category requires different computations than identifying them just as examples of that category (Gauthier et al., 2000). These are so-called within-category object recognition processes. This of course disagrees with the modular theory of functional organization.

Understanding face recognition at the level of neural representation could in fact recruit all of these views and more, as has been shown in previous debates in the Neurosciences. Why is this interesting? Oftentimes mysteries dissapear when a unifying theory has been introduced to explain and account for the seemingly conflicting and divergent evidence. The representation of faces in ventral temporal cortex could appear to be special for three reasons: a specific brain area is used consistently, they may require unique computations to recognize, and that their recognition requires a process which can be selectively impaired, without impacting recognition of other objects.

1.2.1 Selective brain activation?

Converging evidence points to face-specific response in the ventral cortex (Bentin, 2000) (Kanwisher, 2000) (Kanwisher et al., 1997) (McCarthy et al., 1997) (Sergent et al., 1992) (Tsao et al., 2003). Functional imaging studies have reported regions more active during face than nonface viewing by the subjects (Sergent et al., 1992), face versus location matching (Haxby et al., 1994) (Courtney et al., 1997), and faces versus scrambled (distorted) faces (Puce et al., 1995) (Clark et al., 1996). In addition, more activation was recorded in the same area for faces when compared to consonant strings (Puce et al., 1996). All of these studies indicate that parts of the fusiform and inferior temporal gyri of ventral temporal cortex seem to be activated the most when subjects view face images: an area of the brain named the fusiform face area (FFA), located in the right fusiform gyrus.

Interestingly though, the FFA is activated by nonface objects too. A variety of objects elicit more activation of FFA when matched to specific labels rather than abstract ones (ie, 'coelocanth' versus 'fish') (Gauthier et al., 1997). The FFA is also activated during viewing of objects that subjects are experts at identifying, for example, cars or birds when shown to subjects who qualify as car and bird experts (Gauthier et al., 2000). And to further undermine the view that FFA is an innate, face-specific module, people were trained

to be experts at identifying greebles, a fantastic, visually similar set of artificial 'creatures' that the subjects did not see before participating in the experiment. These greeble experts showed consistent activation of the FFA when images of greebles were presented to them (Gauthier et al., 1999). Therefore expertise with objects and identification within-category appear to rely on the same neural representation as face recognition.

1.2.2 Selective Impairment?

Besides the activation studies which point to selective brain activation for faces in humans, there is also work indicating that after brain damage selective impairment to face recognition is possible. What is a double-dissociation?

A double-dissociation is said to exist between two processes A and B if one experimental manipulation affects process A but not B, and in addition there is a different experimental manipulation which affects B but not A. This double-dissociation is said to be sufficient evidence to propose that processes A and B are functionally separable. It is important to stress that double-dissociation is actually possible without assuming that the two processes are separable structurally. That is, we don't have to believe in modules for double-dissociation to imply functional separability (Plaut, 1995).

Prosopagnosia is a condition which impairs face recognition. It is caused by damage to the lingual and fusiform gyri, which contain the FFA (Farah, 2004). If there are subjects for which similar brain damage leads to impaired face recognition who still have intact nonface recognition ability (and vice-versa), this suggests a double-dissociation or functional separation between face and nonface recognition. There is evidence for both subjects that have selectively impaired face recognition with intact nonface recognition (Henke et al., 1998); (Farah et al., 1998); (Farah et al., 1995); (McNeil and Warrington, 1992); (De Renzi, 1986). There are also documented cases of subjects with impaired nonface recognition and mostly

intact face recognition (though this is rare) (Moscovitch et al., 1997). Despite the apparent one-sidedness of the supposed double-dissociation, this double-dissociation has been used to claim that face and nonface recognition are functionally distinct.

However, prosopagnosia in its pure form is nearly non-existent. Patients with lesions to vental temporal cortex that also can't recognize faces well often cannot do within-category classification of non-face objects (Damasio et al., 1982); (Damasio, 1990); (Damasio et al., 1990). This just gives more evidence to the hypothesis that the peculiarities of face recognition are due to general within-category object recognition.

1.2.3 Specialized computation?

Finally, face recognition could require computation of a completely different type than general object recognition. For example, upright faces may be recognized as whole objects rather than as a combination of low-level features (Tanaka and Farah, 1997); (Tanaka and Gauthier, 1997). This means that though feature-based recognition contributes to face recognition, ultimately the whole is greater than the sum of the parts. In addition, spatial relationships between parts of the face (configuration) have been shown to contribute to face recognition (Haig, 1984); (Hosie et al., 1988). This means that our ability to recognize a face is severely impaired if we are just able to see someone's hairline, or nose, or eyes, without having an overall picture of the configuration of the face. Other stimuli, such as faces that have been scrambled or inverted, in addition to inverted houses, can be recognized well by subjects by just viewing their parts (Tanaka and Farah, 1997).

There is also evidence that human subjects are terrible at recognizing inverted faces (Valentine, 1989). A number of electrophysiology studies show that this difficulty shows up as latency delay (ie lag) in brain activation for inverted faces (Bentin, 2000); (Eimer, 1998); (Eimer, 2000); (Rossion et al., 2000). In addition, recognition of faces by prosopagnosics

improves dramatically when they are inverted (Farah et al., 1995). This study then suggests that prosopagnosics are somehow using their intact object recognition process, and that this gives further evidence to the functional dissociation between face and general object recognition.

However, the inversion effect is not unique to faces. It has also been shown to manifest with non-face object experts (Gauthier et al., 1997). Dog-show judges, for example, have more difficulty recognizing inverted dogs (Diamond and Carey, 1986). Car and bird experts also have similar difficulties (Gauthier et al., 2000). Even subjects who were trained to become experts in greeble recognition during the experiment showed inversion effects (Gauthier et al., 1999), and even a dependence on configural cues, but only when the greebles were upright (Gauthier and Tarr, 1997); (Gauthier and Tarr, 2002). So the inversion effect may just be a side-effect of expertise with representing upright objects using spatial configuration (Carey and Diamond, 1977).

It is quite clear that there is something special going on during face recognition. However, it could be something that isn't just unique to faces. Other evidence has pointed to an alternate hypothesis. However this evidence has not indicated one theory which can account for all of the conflicting evidence, which suggests that there is a much more fundamental theory which could explain away the conflict. Notice that none of these studies spoke of the topography or structure of the brain, except in the broadest of generalities. This information, as it happens, may provide the theory we so desperately seek.

1.2.4 Eccentricity bias and object recognition

Other theories

It is important to survey some more results which call other coherent theories into question. For review, the other theories are the face modularity and within-category expertise hypotheses. The first assumes that faces are represented in a domain-specific way in the ventral temporal cortex, with a special brain area, and specialized computation (Kanwisher, 2000). The rest of ventral temporal cortex is assumed to be for general object recognition. The within-category expertise theory suggests that object representations are organized according to the type of computation required (Tarr and Gauthier, 2000). Faces are recognized at the level of individuals, in a highly specific, within-category manner. This is presumably because we become experts in face recognition. For objects which we are unfamiliar with, recognition is done at a very abstract, between-category level, for example, identifying a mexican fender telecaster as a guitar, not an elephant.

With regards to the face module hypothesis, the strong activation of an occipito-temporal object-selective region when subjects view pictures of that object does not mean that this is where all the object recognition happens (Grill-Spector et al., 1998a). At least, not this bit of evidence alone. There is evidence that representations of objects are not restricted to category-specific structurally distinct modules (Ishai et al., 1999). For example, when identifying upright faces and houses, FFA activation is higher during house recognition than vice-versa (Haxby et al., 1999). In fact, there are suggestions that objects are represented in a much more efficient way, distributed topographically across the ventral temporal cortex.

As regards within-category expertise, no mention is made of the topography of ventral temporal cortex. One interesting fact that is unaccounted for by both the modular and expertise hypotheses is that within the posterior fusiform region, selective activation to buildings is always more medial in location than for faces (Epstein and Kanwisher, 1998); (Ishai et al.,

1999). Strictly speaking, we should aspire to a theory which explains the topography of object recognition areas while retaining the strong behavioral consistency of the within-category expertise account of object recognition.

Eccentricity bias

A theory with claims to this ideal is based on the relationship between cortical areas and the organization of early visual areas (Hasson et al., 2002). This theory, proposed by Malach et al. (Malach et al., 2002); (Hasson et al., 2003), argues that the topographic constraints of our early visual system best explain the organization of ventral and dorsal stream visual areas.

Faces are extremely similar as images, as can be seen by histograms of pixel values (see Figures 1.1 and 1.2). This uniformity means that face recognition demands high visual acuity in order to identify small differences in facial configuration unique to individuals. Face recognition also deteriorates rapidly when faces are presented at the periphery of vision, even when corrected for magnification changes (Makela et al., 2001), in order to have the face viewed at the center of the retina. Images of houses, on the other hand, are profoundly non-uniform, as can be seen by their pixel histograms. This suggests that face recognition requires very local, low-level, high-resolution information, while house recognition requires global, high-level, low-resolution information.

In visual cortex there is evidence that parts of the visual field are magnified as a function of their eccentricity (radial distance) from the center and orientation (in degrees, from the horizontal). This retinotopic organization describes well early visual areas and some higher-level visual areas as well. An interesting property of this mapping (which is formally called the log-polar mapping), is that neighboring spots on the retina project to

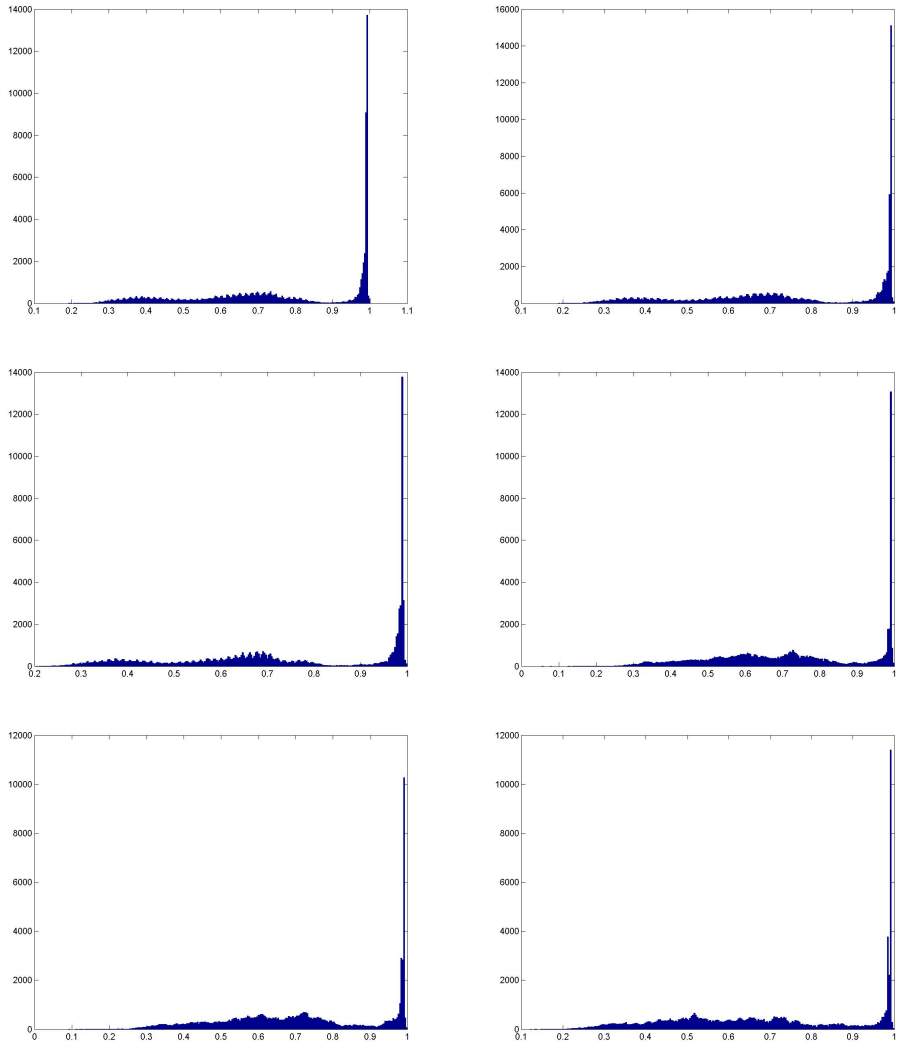

Figure 1.1: These are the histograms with a bin width of 256 of the pixel values in six different images of faces frontally viewed under variable lighting conditions

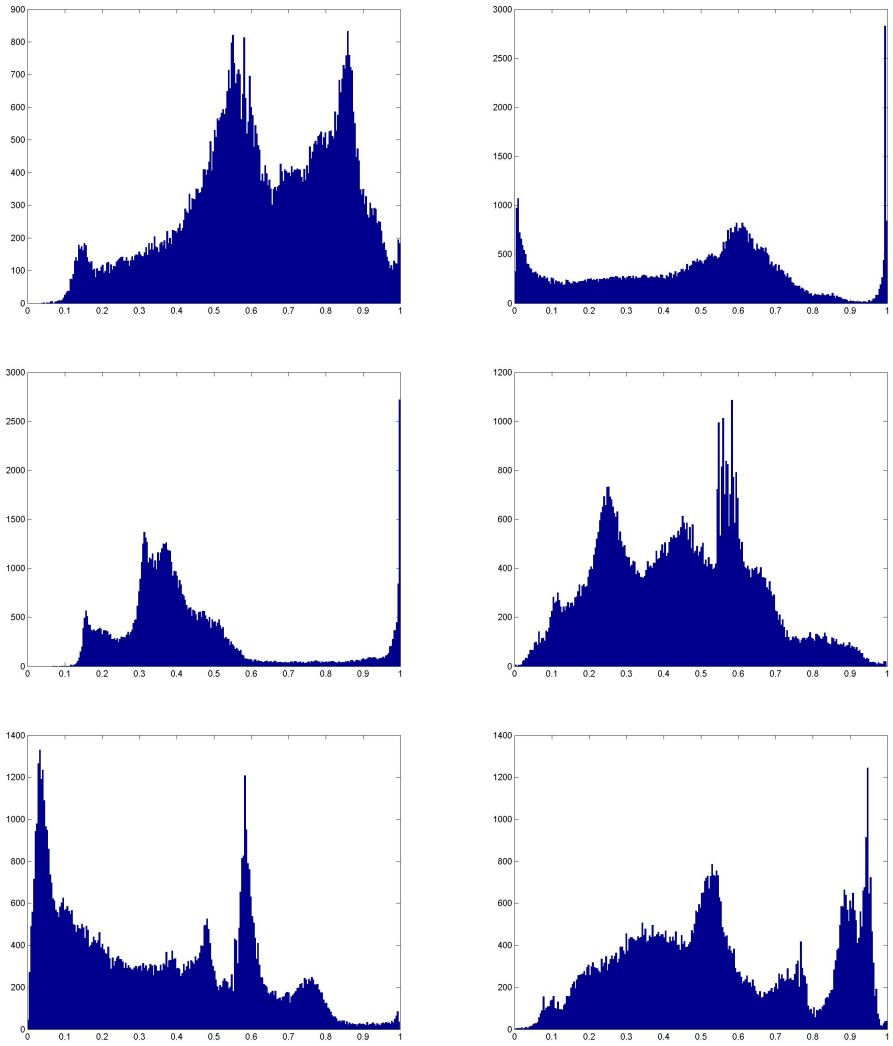

Figure 1.2: These are the histograms with a bin width of 256 of the pixel values in six different images of houses frontally viewed under variable lighting conditions

neighboring cells in the early visual areas. Another interesting aspect of this mapping is that the fovea is over-represented. The center of the visual field has a one-to-one mapping, whereas the periphery has a many-to-one mapping. This means foveal, or more central, projections are higher-resolution. fMRI studies have shown approximately parallel bands of neurons which become active when visual stimuli are presented at different orientations. Perpendicular to these bands are eccentricity bands, which can be viewed as an eccentricity gradient (DeYoe et al., 1996); (Sereno et al., 1995); (Tootell et al., 1985); (Duncan and Boynton, 2003). Because the axes of this representation are orthogonal, it can be approximated by a log-polar transformation.

This eccentricity map may extend into ventral high-order object areas. By using stimuli that tend to activate mid- and high-level visual areas, researchers have shown that these low-level retinotopic constraints are still preserved (Grill-Spector et al., 1998a); (Grill-Spector et al., 1998b); (Hasson et al., 2003); (Levy et al., 2001). To be explicit, the same retinotopic map of central versus peripheral parts of the visual field can be found in regions thought to be involved in face and place perception (Grill-Spector and Malach, 2004). Therefore, the topography of object recognition areas has a very strong eccentricity bias towards the center, which features high-resolution at the center, low-resolution at the periphery, and with distance proportional to distance in the retinal image.

Interestingly, when the location of face-related and building-related regions was projected into an eccentricity map, face images consistently activated center-biased regions, while buildings activated periphery-biased regions, independent of image size (Hasson et al., 2002);(Hasson et al., 2003); (Levy et al., 2001); (Levy et al., 2004); (Malach et al., 2002). Therefore, FFA being more central in the topographic layout of ventral temporal cortex could be due to face recognition requiring high-resolution information (Malach et al., 2002). The preferential bias of upright objects of expertise to a more central eccentricity is because

within-category classification requires high-resolution information and high visual acuity.

It comes as no surprise then that we can throw out our cookie cutters if we adopt this theory. Its elegance is impressive, in that it explains the evidence for both the within-category expertise hypothesis and the face module hypothesis. In addition, it explains the topographic organization of the ventral temporal cortex. It is easy to see why within-category classification requires high-dimensional information about very small variations in images, because objects within the same category are very similar.

Chapter 2

Methods and Algorithms

2.1 Introduction

The problem faced by the nervous system is to represent its interactions with the world as simply as possible but no simpler. The costs of an overly-complex model are more than just aesthetic. A naive approach is to treat the complex interactions with the world as the input to a system which dynamically represents these interactions in a way that is both minimal and goal-directed. The fundamental problem as we have framed it then is to find a representation or model of the input which is both efficient and correct.

2.2 Bayesian Inference

In the specific case that we have inputs and we wish to discover the probability distribution that best reflects the statistics of the input, we can use Bayesian Inference to find the best model to describe the inputs:

$$p(i|m) = \frac{p(m|i)p(i)}{p(m)}$$

This is the so-called posterior distribution, encoding our *a posteriori* (after the fact) knowledge of the input given the model. This posterior distribution is more peaked when our uncertainty about which model m is best for the input is very low. Note that this approach is equally valid using discrete probability distributions as well.

$p(i)$ is a density function which follows directly from our knowledge of the statistics of the input, for example, whether inputs are highly correlated or uncorrelated. This *a priori* (prior) knowledge about the inputs allows us to find a good model more accurately in less time if we have a significant amount of information about the statistics of the input already (strictly, in numbers of iterations). If our knowledge of the input statistics is inaccurate, then this will in fact increase substantially our uncertainty about which model is best and slow convergence.

$p(m|i)$ is the likelihood of the model given the input. This gives us an estimate of how probable a specific model is given the input that we have seen so far.

A specific approach towards finding the best model is to use Bayesian Inference to find a posterior distribution for different models q given the true model t. A very simple way in which this distribution can be used is to find the most probable model q, which corresponds to the model which gives the highest posterior probability $p(q|t)$, the peak of the posterior. This is called the MAP (maximum a posteriori) estimate, is defined as follows:

$$q_{map} = arg\max_{q} (p(q|t)p(t))$$

An often used special case of a MAP estimate is Maximum Likelihood or ML estimation, which is MAP estimation but with a uniform prior distribution on t, so the MAP estimate is equivalent to maximizing the averaged likelihood. The ML estimate converges to the empirical distribution of frequencies actually seen in the data, and assumes no prior knowledge at all about the input. Why it is used often in practice is that it is easy to approximate, with many ML estimates having closed-form solutions.

2.3 Function Approximation

If we wish to model not only the structure of the input, but approximate a function from inputs to behaviors, then we need more information than just the inputs. Specifically, we need an idea of how close we are to the ideal behaviors. The problem of finding a functional mapping from inputs to behaviors that minimizes some Energy or Cost functional is called function approximation.

If our inputs are real-valued but our behaviors must be some sort of category, class, or label,

then the task is called classification. A very relevant example of classification is when we see someone and immediately know their identity based only on their hairline.

2.4 Graphical Models

2.4.1 Random Variables

A random variable X is simply a variable which takes on a specific value $X = x$ with probability $P(X = x)$. A classic example is the random variable signifying the face that comes up when you roll a six-sided die, where each face comes up with a probability $P(X = f) = \frac{1}{6}$.

2.4.2 Statistical Independence

Two random variables A and B are independent under the following conditions:

$$P(A, B) = P(A)P(B)$$

$$P(A|B) = P(A)$$

This is formally denoted as $A \perp B$. What statistical independence intuitively means is that if we know the value of random variable A, we can't predict the value of B and vice-versa. A and B are uncorrelated and, in some interpretations of probability, behave no dependence on each other at all. That is, if A changes, B won't change in a similar way.

A generalization of statistical independence is conditional independence. A is conditionally independent of B given C if

$$P(A, B|C) = P(A|C)P(B|C)$$

$$P(A|B, C) = P(A|C)$$

which is denoted as $A \perp B|C$. The intuitive meaning behind conditional independence is that if we know C, A and B are statistically independent. That is, knowing the value of C makes A and B uncorrelated, unrelated, and thus A and B can be manipulated independently without fear of a change in B causing a change in A, assuming we know the value of C.

2.4.3 Graph Theory Preliminaries

A *walk* is defined as an alternating sequence of edges and vertices, with each edge being incident to the vertices immediately preceding and following it in the sequence. A *trail* is a walk with no repeated edges. A *path* is a walk with no repeated vertices. A *cycle* is a closed trail with at least one edge and with no repeated vertices except that the first vertex is equal to the last vertex.

A directed graph is a graph composed of edges directed from one vertex to another. A directed acyclic graph is a directed graph which contains no cycles.

2.4.4 Graphical Models

A Graphical Model is a graph $G = (V, E)$ with each vertex representing a random variable such that, for each vertex $v \in V$,

$$v \perp nondescendants(v)|Pa(v)$$

that is, each random variable in the graph is conditionally independent of its non-descendants in the graph given its parents. This is the so-called Local Markov Assumption (Pearl, 1988). As a consequence of the local markov assumption, the joint distribution of the random vari-

ables in the graph has the following form:

$$P(v_1, ..., v_n) = \prod_{i=1}^{n} P(v_i | Pa(v_i))$$

So one way to look at a graphical model is as a compact way of representing interacting random variables with known conditional independence properties. In fact, it has been shown that from a graphical model we can obtain a subset of the conditional independences between the random variables and vice-versa. So we can translate any set of assumptions about independence into a graphical model. In fact, using graph theory, we can actually derive more conditional independences, if we consider a special type of graphical model which has directed edges:

2.4.5 D-separation

A trail starting at v_1 and ending at v_k in a directed graphical model is an *active trail* when variables $O \subset v_1, ..., v_n$ are observed (we know their values) if for each consecutive triplet in the trail:

1. $v_{i-1} \rightarrow v_i \rightarrow v_{i+1}$ and v_i is not observed
2. $v_{i-1} \leftarrow v_i \leftarrow v_{i+1}$ and v_i is not observed
3. $v_{i-1} \leftarrow v_i \rightarrow v_{i+1}$ and v_i is not observed
4. $v_{i-1} \rightarrow v_i \leftarrow v_{i+1}$ and v_i or one of its descendants is observed

Variables v_i and v_j are conditionally independent given $Z \subset X_1, ..., X_n$ if there is no *active trail* between v_i and v_j when variables Z are observed. In this case, v_i and v_j are said to be *d-separated* and $v_i \perp v_j | Z$ (Jordan, 1999).

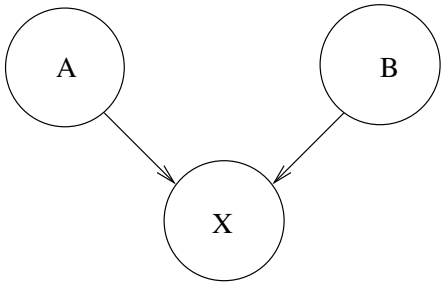

Figure 2.1: Random variable X has parents A and B in this graphical model. Explaining away effects can be seen when we know the probability of X taking on a value. In this situation knowing A or B alone explains away any influence of the other variable.

2.4.6 Explaining away effects

Suppose we have a random variable X with two parents A and B that are not connected to each other in a graphical model. Explaining away means formally that $P(A = a|X = x, B = b) \leq P(A = a|X = x)$, which means that if we observe A, it is sufficient to explain any changes observed in X (Pearl, 1988). The same holds for B as well, because either one alone will explain changes in X. This generalizes to more than two parents as well. This graphical model is shown in Figure 2.1.

2.4.7 Hierarchical Bayesian methods and Graphical Models

If we have some very general idea of how one variable varies with respect to another, we say that one is proportional to the other in some way, multiplied by a constant of proportionality. Usually the constant is determined empirically by trial and error. However, it is actually possible to learn what this constant should equal, giving a more principled approach to learning a model of the relationship. This constant of proportionality is a type of parameter, which we tune to change the relationship between two variables. Let's say we have two random variables, P, which stands for the presidential candidate who wins the election, and IQ, which is either 1 or 0, depending on whether a candidate is intelligent or not. We would hope that these two variables are related, so we say $P(P = p) = \xi P(IQ = 1|P = p)$. Un-

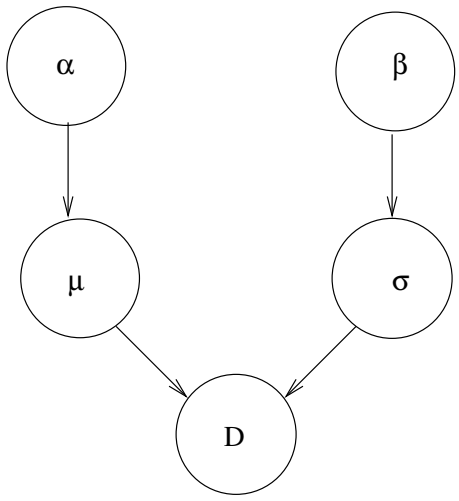

Figure 2.2: This is a graphical model for hierarchical bayesian estimation of the mean and standard deviation of a gaussian fitted to data given prior information about the mean and standard deviation.

fortunately for us, ξ ended up being a negative number for these last two elections. Maybe if we had some way of predicting this, many things would have changed. Once again, bayesian inference has a way to this, by treating ξ as a random variable and calculating a probability for it taking on different values (Good, 1987); (Good, 1965).

A more orthodox presentation of so-called hierarchical bayesian inference is when we wish to fit a gaussian distribution to some new data D for something we've already seen before. This prior data is very extensive, and we have found from it a specific value for the mean and variance (α and β, say). But we can't say that the mean and variance will be the same for this new data. At the same time, we can use what we know already by treating μ and σ as random variables and assigning a high probability to them taking the values we found from the prior data. This is called the empirical bayesian approach. The subjectivist approach to hierarchical bayesian inference is similar, but derives the prior from formal assumptions about the model. A graphical model for this situation is shown in Figure 2.2.

2.5 Neural Network Models

A neural network is a compact representation of a function mixing linear algebra, statistics, and signal processing (Bishop, 1996). Each "node" in a neural network has a set of input connections and a set of output connections. The type of neural network which will be studied here is a classic three-layer neural network, defined by these equations:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

$$y_h = g(\sum_i x_i w_{ih}) \tag{2.2}$$

$$y_o = g(\sum_h y_h w_{ho}) \tag{2.3}$$

Where there are three layers: the input, hidden, and output layers, and with y being the output of a single unit in a specific layer. $g(x)$ is a transfer function applied onto the weighted sum of the inputs to a node.

From a practical perspective, the weights (indicated by w) must be learned so as to approximate some function. In the case we are interested in the mapping from real-numbered input to real-numbered outputs. To learn a mapping, we must find the weights which minimize some error function or loss function. In the classic 3-layer supervised learning task, we are provided not only with the inputs, but also target or desired values d_o for the outputs by a teacher. The error functional which we are interested in then is:

$$E_D = \frac{1}{2} \sum_o (d_o - y_o)^2$$

This error or loss function is called sum-squared error or quadratic loss, which is appropriate when we assume that there is fixed additive gaussian noise added to the outputs of the networks (and wish to maximize the likelihood of the choice of weights given the data) (Duda et al., 2001).

2.6 Topographic Bias

The theory and justification for Topographic Bias is to provide a mechanism for favouring short connections and enforcing locality of representations in the hidden layer. This constraint is inspired by brain organisation, which must permit sufficient connectivity among neurons to carry out the necessary computations while still fitting the total axonal volume in the skull (Jacobs and Jordan, 1992). In the case of the brain, this constraint is drastic, if the brain's approximately 10^{11} neurons were placed on a sphere and interconnected with 0.1 m radius axons, fitting the total axonal volume in the skull would require a sphere over 20 km in diameter (Nelson and Bower, 1990). Also, another constraint on neural organization is the use of metabolic energy. Only a limited amount of metabolic energy can be used by the brain, and the neural organization suggested to optimally conserve energy is one that has tightly-interconnected clusters of neurons which have sparse projections to other clusters (Laughlin and Sejnowski, 2003). Topographic bias, in a formal sense, is implemented by scaling the magnitude of the error derivatives by a gaussian function of the length of the connection. Formally, the partial derivative of the error with respect to $\mathbf{w_{ij}}$ is multiplied by the following:

$$\tau(l, j) = e^{\frac{-l(i,j)^2}{\sigma_\tau^2}}$$

where

$$l(i, j) = \sqrt{(j - i)^2}$$

This can be simplified to

$$\tau(i, j) = e^{\frac{-(j-i)^2}{\sigma_\tau^2}}$$

2.7 Bayesian analysis of Neural Networks

As a result of the nonlinearity introduced in the hidden layer, analysis of what computations are being performed by the classic three-layer network are often intractable. However, it is

possible to analyze the computational properties of the network in more detail if we adopt the Bayesian analysis framework ((MacKay, 1992)). Similar to the proof/program duality in the field of programming languages, not only can this analysis framework give us a way to understand the computations of the network, but also create more general models from this understanding. Or, put another way, we can compile a general model for optimal performance, which can generate new learning algorithms and a common framework to understand all of them.

An ingenious observation makes analysis of the classic three-layer network more tractable: the two layers of weights can be analyzed as part of a general weight matrix that is being learned. This new weight matrix can be written as $\mathbf{w_{io}}$, and is constrained by the dimensions of smallest layer. Similarly, the biases at each layer can be combined with the weights to form the parameter $\mathbf{w}$. The learning of the weights then becomes a very simple problem: how do we find the optimal $\mathbf{w}$ to map from the inputs to the outputs? Because we have effectively combined the intermediate layers and collapsed the three-layer network into a two-layer network, the energy functional changes slightly as well:

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_h \sum_o (d_o^{(h)} - y_o(x^{(h)}; \mathbf{w}))^2$$

We wish to find the weights which minimize this energy functional (the error) (MacKay, 1992). The interesting parallel to statistics with this objective is that it is equivalent to maximizing the likelihood of the weights assuming that the desired values have a fixed level of gaussian additive noise applied to them. To see this, we note that the objective is to find the best $\mathbf{w}$ possible, given the inputs to the network, the network architecture, and some level of noise we are assuming exists in the desired values. D is the data fed to us as input, A is the network architecture, and β is the amount of noise in the desired values. According to our discussion of Bayesian inference, this yields the following rule for inference of the

best $\mathbf{w}$:

$$P(\mathbf{w}|D, \beta, A) = \frac{P(D|\mathbf{w}, \beta, A)P(\mathbf{w}|A)P(A)}{Z_m(\beta)}$$

where Z is the partition function. If we assume that all weight vectors are equally likely and so are all possible architectures, then this simplifies to just the normalized likelihood. Due to the gaussian noise added to the desired values, the likelihood can be defined as follows:

$$P(D|\mathbf{w}, \beta, A) = \frac{e^{-\beta E_D(\mathbf{w})}}{Z_D(\beta)}$$

It is easy to see then that under the uniformity assumptions we made the log-transformed posterior is just the average log-likelihood, which is

$$\log P(D|\mathbf{w}, \beta, A) = -\beta E_D(\mathbf{w})$$

Minimizing the sum-squared error defined above is thus equivalent to maximizing the log-likelihood (MacKay, 1992). This of course is a more abstract view of what learning the weights must do. We introduce and justify gradient descent to see how the weights can be learned over time from the input. Then we will return to Bayesian analysis of the three-layered network after discussing rigorously energy functionals.

2.8 Stochastic Gradient Descent

To maximize the log-likelihood and minimize the error, we introduce gradient descent, which is by far the most popular algorithm for optimizations of this sort. The basic idea is to take the partial derivative of the Energy functional with respect to the variable being optimized and update the variable additively over every iteration, weighted by a learning rate (the sign of the additive change determines whether it is maximization or minimization).

The partial derivative of interest is clearly

$$\frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}}$$

because we wish to minimize the error of our weights over time. Our weight change to minimize this partial derivative is then

$$\mathbf{\Delta w} = -\epsilon \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}}$$

In our prior analysis, the quantity of interest was the log-likelihood, which if we take the partial derivative yields the following:

$$\log P(D|\mathbf{w}, \beta, A) = -\beta E_D(\mathbf{w})$$

$$\frac{\partial \log P(D|\mathbf{w}, \beta, A)}{\partial \mathbf{w}} = -\beta \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}}$$

It is easy to see that using gradient descent on the error is equivalent to using gradient ascent on the log-likelihood, and, notably, the learning rate is actually equal to β, the amount of noise that is assumed to be in the desired values given to the learner. Backpropagation can be understood as gradient descent on the error in $\mathbf{w}$-space (MacKay, 1992).

2.9 Energy Functionals: Proof/Model Isomorphism

In the words of Schrodinger, "There is, essentially, only one problem in statistical thermodynamics: ...to determine the distribution of an assembly of N identical systems over the possible states in which this assembly can find itself, given that the energy of the assembly is a constant E." He then goes one to say that the individual states are independent and have energies that sum to the constant E. Finally, he defines E as a constant of motion useful for studying a system (Schrodinger, 1989).

John Hopfield first noted the similarity between memory and minimizing an energy functional (Hopfield, 1982), and the mapping between statistical thermodynamics and learning algorithms has been fruitful, yielding the EM algorithm (Dempster et al., 1977) and Boltzmann Machine (Ackley et al., 1985) among others. However, the sense in which energy functionals are used in that work is much more specific, tied to the existence of a Lyapunov functional (Hopfield, 1982). We wish to study energy functionals with an interpretation closer to Schrodinger's.

There is a very interesting link between the log-posterior and the minimization of some energy functional. This link was implicit in our prior analysis of the three-layer network. Here we make it explicit. If our energy functional E is to be minimized with respect to θ, then this immediately yields a probability model for the posterior, based on the Boltzmann distribution (Schrodinger, 1989):

Theorem 1. *For any functional E quantifying cost, energy, error, or loss associated with believing parameters θ, when approximating random variable D, assuming a variability g in the accuracy of our estimate:*

$$P(D|\theta, g) = \frac{\delta(\theta)e^{-gE(\theta)}}{Z_E(\theta)}$$

where $Z_E(\theta) = \sum_\theta e^{-gE(\theta)}$ is the partition function and $\delta(\theta)$ is the degeneracy function which gives the number of other parameters which have the same energy $E(\theta)$ as θ.

Proof. Consider an assembly of N parameters θ, each with an energy $E(\theta) = \epsilon_i$. We assume that the most probable distribution of parameters among energy states is a maximum of W, the number of different assignments. Each microstate g_i has energy ϵ_i, and the number of ways a parameter could have energy ϵ_i is g_i. If we have n independent parameters, the number of ways the parameters can have energy ϵ_i is $(g_i)^n$.

The number of ways N parameters can be distributed among the possible energies is $\prod_i (g_i)^{n_i}$, where n_i is the number of parameters with each energy, subject to $\sum_i n_i = N$ and $\sum_i \epsilon_i = E$. Since there are N parameters and n_i of them fit in the i^{th} microstate g_i, the total number of ways to distribute all parameters into the microstates is:

$$W = \frac{N!}{\prod n_i!} \prod g_i^{n_i}$$

If N is large, stirling's formula($\log n! = n \log n - n$) can be used. We take the maximum of $\log W$, subject to constraints to take advantage of stirling's formula. This can be done by using the Lagrange multipliers implied by our constraints:

$$f(n_i) = \log W + \alpha(N - \sum_i n_i) + \beta(E - \sum_i n_i \epsilon_i)$$

When we use Stirling's formula and set $f(n_i) = 0$ to find the maximum, we obtain

$$n_i = \frac{g_i}{e^{\alpha + \beta \epsilon_i}}$$

Since this is just an approximation using combinatorics, we observe that the joint probability of entering a set of microstates g_i is multinomial, and that the only constraint that isn't satisfied by our expression for W to be a multinomial distribution is that g_i isn't normalized. We assume that the a priori probability of entering a microstate is uniform. Instead of normalizing g_i, we instead find the maximum likelihood estimate for the probability of entering the microstate g_i, assuming that the probability of a particular assignment of energies to microstates is multinomial:

$$P_{ml}(g_i) = \frac{n_i}{N}$$

is the maximum likelihood estimate for the probability of entering a microstate if the joint

distribution on microstates is multinomial. Also, remember that $\sum_i n_i = N$. Therefore,

$$P_{ml}(g_i) = \frac{\frac{g_i}{e^{\alpha+\beta\epsilon_i}}}{\sum_i \frac{g_i}{e^{\alpha+\beta\epsilon_i}}}$$

$$P_{ml}(g_i) = \frac{g_i}{e^{\alpha+\beta\epsilon_i} \sum_i \frac{g_i}{e^{\alpha+\beta\epsilon_i}}}$$

$$P_{ml}(g_i) = \frac{g_i e^{-(\alpha+\beta\epsilon_i)}}{\sum_i \frac{g_i}{e^{\alpha+\beta\epsilon_i}}}$$

$$P_{ml}(g_i) = \frac{g_i e^{-(\alpha+\beta\epsilon_i)}}{Z_{\alpha,\beta}}$$

If we further assume that $\alpha = 0$ and $\beta = g$, and in addition, remember that $\epsilon_i = E(\theta)$, then we have:

$$P_{ml}(g_i) = \frac{g_i e^{-(gE(\theta))}}{Z_E(\theta)}$$

If we take the likelihood of the set of parameters θ and g to be equal to the (maximum likelihood estimate of) probability of microstate g_i, then we have:

$$P(D|\theta, g) = \frac{\delta(\theta) e^{-(gE(\theta))}}{Z_E(\theta)}$$

where $\delta(\theta)$ is simply a function which, given a set of parameters, returns the microstate g_i corresponding to it. This can be understood as the degeneracy of the energy functional $E(\theta)$, because if it is greater than one, then many different parameters end up having the same energy. This of course makes finding a minimum or maximum of $E(\theta)$ harder.

An important special case is when each parameter has a unique energy associated with it, in which case $\delta(\theta) = 1$. □

the log-posterior is simply:

$$log(P(\theta|D, g)) \propto -gE(\theta) + \log P(\theta) + \log \delta(\theta)$$

which if each set of parameters θ is assumed to be equally likely a priori and to have a unique energy, equals:

$$log(P(\theta|D,g)) \propto -gE(\theta)$$

This is Mumford's definition of an energy functional, and the common one which is used in the computer vision community (Mumford, 1994). Therefore, if we define an energy functional, we automatically have a statistical model of the network. This suggests that the output of an energy functional can be viewed as a random variable.

Note that this model assumes a priori that all sets of parameters have a unique energy associated with them. If this is not true, then each term in the probability should be weighted by a different degeneracy factor g_i, corresponding to the number of different sets of parameters for which $E(\theta)$ is identical. So the common definition of energy functional as it is used in both the neural networks and computer vision assumes many things statistically. Degeneracy is actually quite common. In fact, the sum-squared error or quadratic loss has a degeneracy $\delta(\theta) = 2$ for all θ because of the reflective symmetry of a quadratic function. A proof of this is not relevant to the current work, because we never treat degeneracy directly. However, in another paper that is in progress it is calculated in full from the symmetries of the energy functional (and thus invokes group theory). In future work it is likely that degeneracy in energy functionals can be used to find optimal energy or loss functionals for a particular learning problem. This in fact follows from the fact that the log-likelihood of the probability associated with the energy function includes a penalty term for degeneracy.

Curiously, in the case of statistical mechanics, this assumption was made exclusively by Boltzmann and other proponents of classical physics (kinetic theory). This alternative form, which keeps the degeneracies, is attributed to a reformulation of the Boltzmann distribution in the early history of quantum mechanics (Schrodinger, 1989).

This statistical model is also equivalent to assuming that the noise model is fixed and gaussian (with a mean of zero and standard deviation $\frac{1}{\sqrt{2g}}$), but only if the energy is in some way a quadratic function. Thus an energy functional, which tells us a great deal about the functioning and learning of the system, also provides us with a statistical model of that system.

Formally, the statistical model (Boltzmann distribution) has the property that low-energy states are more probable than other states when the system is at equilibrium and has converged to a solution (Schrodinger, 1989).

The following sections assume that $\delta(\theta) = 1$ for all θ. This is because gradient descent as it is used in backpropagation assumes it and that is the algorithm we are analyzing.

2.9.1 Composite Energy Functionals

Another intriguing bit of analysis is what happens when energy functionals are summed. Specifically, what does this do to the statistical model of our network? If there are two constraints at play in the network, defined by energy functionals E_P and E_Q, then their composite energy functional is simply $E_{PQ} = E_P + E_Q$. By Theorem 1, the following is true of the composite functional E_{PQ}:

$$P(D|\theta, g) = \frac{e^{-g_{PQ}E_{PQ}(\theta)}}{Z_E(\theta)}$$

$$P(D|\theta, g) = \frac{e^{-(g_{PQ}/g_P)E_P(\theta)}e^{-(g_{PQ}/g_Q)E_Q(\theta)}}{Z_E(\theta)}$$

$$P(D|\theta, g) = \frac{e^{-(g_{PQ}/g_P)E_P(\theta)}e^{-(g_{PQ}/g_Q)E_Q(\theta)}}{Z_{E_P}(\theta)Z_{E_Q}(\theta)}$$

$$P(D|\theta, g) = \left[\frac{e^{-(g_{PQ}/g_P)E_P(\theta)}}{Z_{E_P}(\theta)}\right]\left[\frac{e^{-(g_{PQ}/g_Q)E_Q(\theta)}}{Z_{E_Q}(\theta)}\right]$$

The same, of course, can be said of any composite energy functional. A composite energy functional represents the log-posterior of multiple independent gaussians. That is, if the outputs of both E_P and E_Q are random variables, the probability of both having a certain value during optimization of E_{PQ} is decomposable: the sum of the probabilities of each.

This means there is a principled way of decomposing energy functionals into basic, statistically-independent energy functionals. This is isomorphic to the statistical mechanics conception of weakly-interacting or independent systems (Schrodinger, 1989): the network is doing learning of multiple things at once, rather than learning one thing. This is equivalent to the concept in optimization of a "tradeoff": learning with respect to one measure alone prevents learning anything else. The best learning algorithm for a composite energy functional will try to learn best with respect to all measures. In other words, $E_P \perp E_Q$. However, when E_{PQ} is observed, then the two become dependent. This is shown perfectly in the following graphical model (Figure 2.3):

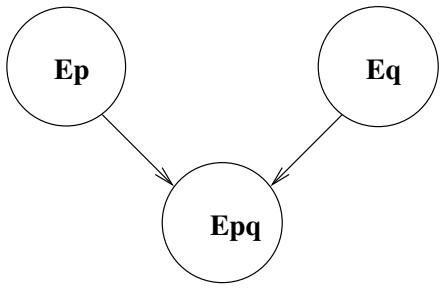

Figure 2.3: This is a graphical model for a composite energy functional E_{PQ}

Interestingly, when E_{PQ} is observed, along with E_P, the probability of E_Q goes down. This is a classic example of the explaining away effect mentioned before. In some sense then, E_P and E_Q are inversely related, when the total energy is known, and increases

in certainty about E_P mean more uncertainty about E_Q. This is another, more formal way of analyzing a tradeoff in optimization.

2.9.2 Basic Energy Functionals

A basic energy functional is an energy functional $E(\theta)$ that cannot be simplified into statistically-independent components during optimization. That is, when learning using this energy functional, only one thing is being learned or optimized. All terms in the energy functional are coupled, or strongly related. A concrete example of such an energy functional is $E_B = E_P E_Q$. By Theorem 1 the statistical model for $E_B(\theta)$ is:

$$P(D|\theta, g) = \frac{e^{-g_B E_B(\theta)}}{Z_E(\theta)}$$

$$P(D|\theta, g) = \frac{e^{-g_B (E_P(\theta) E_Q(\theta))}}{Z_E(\theta)}$$

$$P(D|\theta, g) = \frac{e^{-g_B (E_P(\theta)^{E_Q(\theta)})}}{Z_E(\theta)}$$

which doesn't decompose into two statistically-independent statistical models. That is, if the outputs of both E_P and E_Q are random variables, the probability of both having a certain value during optimization of E_B is not decomposable: both are strongly coupled. This is illustrated in the graphical model for the energy functionals (Figure 2.4):

Interestingly, if any of the energy functionals are known with any certainty, the other two become uncertain (again by explaining away effects). Under no circumstances can any of the energy functionals be independent.

2.9.3 Improper Energy Functionals

When using energy functionals to guide learning, it is often taken for granted that all parts of the energy functional will affect learning. However, this isn't true, especially for what are called improper energy functionals. The definition of an improper energy functional E is

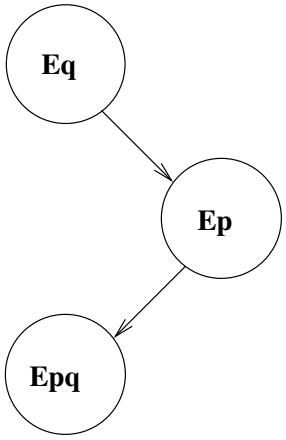

Figure 2.4: This is a graphical model for a basic energy functional E_{PQ}

an energy functional which has terms not being optimized by the learner. In practice, these can result from aspects of a statistical model which come from prior information on the architecture of a network, or other sources. Those terms which are not in terms of variables being optimized by the learner effectively dissapear in the partial derivative with respect to the optimized variables during gradient descent.

2.9.4 An Energy Functional for Topographic Bias

Now we can demonstrate how this expanded set of methods for analysis of energy functionals can be of use. To formally study what impact some constraint to the network has on learning, it is useful to find the partial derivative of the energy functional with respect to the weights. First, we write an expression for the partial derivative of the energy functional E_τ as we defined topographic bias:

$$\frac{\partial E_\tau}{\partial \mathbf{w_{ij}}} = \tau(i,j)\frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w_{ij}}}$$

An energy functional which has this partial derivative is:

$$E_\tau(\mathbf{w}) = \sum_i^m \sum_j^n \tau(i,j) E_D(\mathbf{w_{ij}})$$

with $\mathbf{w}$ being an nxm matrix, and where $\tau(i,j) = e^{\frac{-(i-j)^2}{2\sigma_\tau^2}}$ is a matrix.

We show that the partial derivative is as desired below:

$$\frac{\partial E_\tau}{\partial \mathbf{w_{ij}}} = \frac{\partial(\sum_i^m \sum_j^n \tau(i,j) E_D(\mathbf{w}))}{\partial \mathbf{w_{ij}}}$$

$$\frac{\partial E_\tau}{\partial \mathbf{w_{ij}}} = \frac{\partial(\tau(i,j) E_D(\mathbf{w}))}{\partial \mathbf{w_{ij}}}$$

$$\frac{\partial E_\tau}{\partial \mathbf{w_{ij}}} = \tau(i,j) \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w_{ij}}}$$

an approximate form for E_τ is derived below:

$$E_\tau(\mathbf{w}) = \sum_i^m \sum_j^n \tau(i,j) E_D(\mathbf{w})$$

$$E_\tau(\mathbf{w}) = E_D(\mathbf{w}) \sum_i^m \sum_j^n e^{\frac{-(i-j)^2}{2\sigma_\tau^2}}$$

$$E_\tau(\mathbf{w}) = \frac{E_D(\mathbf{w})}{e^{\sigma_\tau^2 - 2}} \sum_i^m \sum_j^n e^{i^2 + ij + j^2}$$

Therefore, the standard deviation of the gaussian we are scaling the weights by (σ_τ) should be as high as possible, but no higher. That is, this energy functional penalizes decreases in the standard deviation of the topographic bias. However, if the standard deviation is too high, then the resultant energy will be low, even if the error of the current set of weights is high. Therefore the standard deviation should also avoid being too high. This tradeoff between being able to correct for changes in error and not throw away too much global information is best resolved with a fairly localized, small standard deviation for the to-

pographic bias (the exponential grows quickly). Also, there is a very strong penalty on increases of mn, the size of the weight matrix. For a fixed topographic bias and network architecture, this energy linearly increases with increases in error.

The statistics of the model are interesting chiefly because in the context of the statistical model implied by E_τ (by Theorem 1), the energy formed from the sum-squared error is a basic energy functional which cannot be decomposed into independent gaussians. That is, E_τ cannot be decomposed into a combination of simpler energy functionals, and what this means is that the error isn't a separate term in the model. We can only discern the error from E_τ, which means that everything else but $E_D(\mathbf{w})$ must remain fixed, or else the likelihood of the weights will not be related closely to their error. This of course also means that we have to choose the architecture of the network and the amount of topographic bias very well before even seeing any inputs, which can only happen if we have a prior model of both these things.

2.9.5 An Energy functional for weight decay

The energy functional for weight decay is

$$E_{\mathbf{w},\alpha} = \frac{1}{2} \sum_i \sum_j \alpha \mathbf{w}_{ij}^2$$

which penalizes weights that grow too large. In a sense, minimizing this energy functional means using models that are as simple as possible. Also, and most importantly, learning the weights to fit the data precisely makes the neural network inflexible. Specifically, if input that the network has never seen before is presented to it, it does horribly. The idea is that the network should learn weights that will generalize to new inputs by keeping the weights as small as possible (MacKay, 1992). Intuitively, this means that an overly complex model which has an internal model of every type of change in the input leads to what statisticians

call "overfitting," the equivalent of fitting a function to plotted data by connecting the dots in the scatterplot. Obviously such a function would fail to fit a new datapoint unless it was identical to one already seen before.

2.10 Bayesian Ideal Observer Analysis in psychophysics

The central tenet of ideal observer analysis is a constructed "ideal observer," which is a theoretical device that performs a given task in an optimal way given available information and specific constraints. Ideal observers do not perform without error, but at the physical limit of what is possible given the constraints imposed and information available. Because there is an inherent complexity and uncertainty in the visual environment, the ideal observer model must be described in probabilistic (statistical) terms (Geisler, 2003). This type of analysis has been successfully applied to much of lower-level visual perception, including characterizing the simple-cell responses in V1 (Simoncelli and Olshausen, 2001) and multi-sensory integration (Knill and Richards, 1996).

Ideal observers are derived from Bayesian statistical decision theory. In this framework, complex goals are represented with a utility or loss function, $u(\mathbf{r}, \omega)$, which specifies the cost associated with making response r when the state of the environment is ω (Berger, 1993). The optimal response for an ideal observer is then the MAP estimate over all states:

$$\mathbf{R} = arg\max_{\mathbf{r}} \left[\sum_{\omega} u(\mathbf{r}, \omega) p(\mathbf{S}|\omega) p(\omega) \right]$$

and before going too astray from our original formulation of learning problems, it is assumed that the particular utility function $u(\mathbf{r}, \omega) = -g_{\mathbf{r}} E(\mathbf{r})$ can be chosen. Finally, this is only true for the retinal image as it is presented. To make ideal observer analysis amenable to constraints, Geisler et al. suggest transforming the stimulus S into an intermediate representation Z and using the same formula for an optimal response $\mathbf{r}$, replacing S with Z

(Geisler, 2003). In addition, the utility, loss, and, ultimately, energy function can be modified to introduce constraints. So the final modified expression for the response of an ideal observer is:

$$\mathbf{R} = arg\max_{\mathbf{r}} \left[\sum_{\omega} -g_{\mathbf{r}} E(\mathbf{r}) p(\mathbf{Z}|\omega) p(\omega) \right]$$

2.11 Ideal Classifier Analysis

In the case of an ideal observer in which there are c_n different classes, and the observer must respond with the correct class c_i given a stimulus S or intermediate representation Z, which is a special case of the general ideal observer. The possible states of the environment are the classes $\mathbf{c} = c_1, ..., c_n$, the responses are the class labels, and the utility for each response is either 1 or 0. This yields the following ideal classifier:

$$\mathbf{C} = arg\max_{\mathbf{c}} \left[\sum_{\mathbf{c}} p(\mathbf{Z}|\mathbf{c}) p(\mathbf{c}) \right]$$

of course, any classifier can be constructed from a function approximator which tries to find a function from the intermediate representation Z to the class labels c_i that minimizes the energy $g_c E_c$. Thus the ideal classifier can be understood as finding the MAP estimate of the class given the intermediate stimulus Z. The MAP estimate can be learned by a three-layer neural network that learns a mapping $\mathbf{w}$ from the intermediate representation Z to the class labels (learning the MAP estimate for $\mathbf{w}$ allows the network to generate C) (Duda et al., 2001). This means that a particular implementation of the ideal classifier is gradient descent on the energy $g_c E_c$ of the mapping $\mathbf{w}$ from intermediate representation Z to class label c, which if we assume *a priori* that all classes are equally likely is the maximum likelihood estimate, given by:

$$\mathbf{W} = arg\max_{\mathbf{w}} \left(\log P(Z|\mathbf{w}, \beta, A) \right)$$

$$\mathbf{W} = arg\min_{\mathbf{w}} \left(\beta E_D(\mathbf{w}) \right)$$

2.12 Cognitive Neuropsychology and Computational Models

The logic of double-dissociation has already been discussed in part 1. Most interesting is how to interpret a demonstrated double-dissociation in a neural network. Or, more rigorously, when such an empirically-derived double-dissociation can provide evidence that the computational assumptions behind a model are valid. Plaut and colleagues have discussed this at great length (Plaut, 1995); (Plaut and Shallice, 1993). The most important criteria is to know what assumptions are being made by the model, and justify each of those assumptions. The early attempts at computational models of patients with brain damage were largely unprincipled in this regard, notably the early Hinton and Shallice model (Hinton and Shallice, 1991).

One aspect of using neural networks as ideal models of a task is that there are many fewer units in a neural network than in an actual network of neurons in the brain. Thus each unit cannot be taken to be equivalent to a real neuron when interpreting the behavior of the ideal observer. Instead, each unit should be interpreted as representing a large cluster of neurons. That is, ideal observer models using neural networks must be analyzed as a coarse measure of the impact of structure on behavior. When a single unit is lesioned in our ideal observer, it is equivalent to lesioning a large number of interconnected neurons (Plaut and Shallice, 1993).

2.13 Eccentricity Bias Models of Object Representation

2.13.1 Ideal in Classification Error

In the ideal observer that was constructed, log-polar transformed images of faces and houses at varying scales were the inputs Z. These transformed stimuli are meant to stand for the intermediate representation Z of retinotopic visual cortex. The responses available are 34 distinct faces and 9 distinct houses. The desired outputs specified by the trainer are the cor-

rect identities corresponding to each input. To be specific, a feedforward, three-layer neural network was taken to be the ideal observer for this task whose weights were learned using backpropagation. The inputs are bitmaps, with each unit in the input layer corresponding to a pixel in the bitmap (1024 units total). The hidden layer is a bottleneck, significantly smaller than the input layer, and a topographic bias was imposed on the hidden layer weights in order to enforce locality. The number of hidden units is large initially, but progressively diminished as the ideal observer is corrupted. The hidden layer receives input from every input unit. A small weight decay term is used to allow the network to learn more general patterns, and a negative biasunit was added to both layers. The output layer is localist and corresponds to each of the different responses. Due to the log-polar transformation of the images, increasing scale means horizontal translation. This means that face images will activate primarily center-biased portions of the hidden layer, with house images extending more peripherally in the hidden layer. The weights between the input units and the hiddens, in addition to the layer between the hidden units and the outputs are initially random values in the range $[-0.25, 0.25]$ before training. The weights from bias units are initially set to random values close to zero. The metric used for topographic bias is the horizontal distance between two units. Note that delta-bar-delta and momentum were used, but since they only speed up the learning, their effects on learning (ie on the energy functional) is ignored. See Figure 2.5 for a picture of the initial network architecture, including an example of the log-polar face input.

The face inputs are 32x32 bitmaps, which vary in the spacing between the eyes, nose, and mouth. The building inputs were different in a more global sense, with changes such as bigger windows, a different base, and different shapes for the roof. In addition, for the training regime, each input was presented at nine different scales, decreasing from 100 to 60 percent in 5 percent decrements. If the network was presented with the pure bitmaps, it could possibly learn the within-category classification task based only on the number of active units in

the input layer. To circumvent this, we used gaussian smoothing, and presented the blurred image to the ideal observer. Also as a part of the training, smaller face images and larger house images were presented more frequently as input to the ideal observer. Specifically, the frequency of presentation was approximated by scaling the error derivatives such that as the scale of face inputs increased, their impact on learning decreased. As the scale of building inputs increased, their impact on learning increased. This allows us to impose a prior constraint on the scale of faces and houses inspired by the statistics of actual images (faces are more likely to be small in scale than houses).

In the training regime there is much to be justified. First, there is controversy surrounding the use of backpropagation in computational models of neural systems (Plaut and Shallice, 1993). We are using backpropagation with no claims to the algorithm being implemented in learning representations of objects. Instead we claim that it is an ideal observer for the task of face and object recognition within-category. This is because it is ideal in terms of mean-squared error. The controversy is still alive because backpropagation is a supervised-learning method, which assumes that there is a teacher giving the network explicit feedback. This is well-known to be biologically-infeasible as a model. However, as we shall see later, the assumption that we have access to the correct response for every input can be relaxed without changing the important computational assumptions of the ideal observer's energy functional. The feedforward nature of the architecture is not meant to be any serious claim about the connectivity of the brain, as it is known to be rife with feedback loops. Again, this is considered to be an ideal observer, and thus behaviorally isomorphic to the neural representation of faces (not structurally isomorphic, in detail). Finally, the initially strong negative bias is to speed up learning, because the network will learn to turn off all but one unit anyway due to the localist nature of the output layer.

To analyze the learning of the network, we characterize the energy functional minimized

during learning. The utility of the isomorphism between energy functionals and statistical models is that we can translate between the two. This means that if we have a statistical characterization of the network's learning, we can convert that into an energy functional, and vice-versa. We will take advantage of that with analysis of this network. Part of what we know already about the learning behavior is in the language of energy functionals, and the other part is in the language of statistics. We know that the energy functional that impacts learning directly and is optimized is the sum of $E_W(\mathbf{w})$ and $E_\tau(\mathbf{w})$, because of the imposition of a topographic bias and weight decay during each weight update. We also know that the architecture of the network in the hidden layer has face units closer to the inner (leftward) section of the output layer, whereas the house output units are in the outer section of the hidden layer (as the log-polar transformed image is being learned). We can model this as two gaussian priors for the network architecture, one for faces, and one for houses, with the mean of the face prior being 0, and the mean for the house prior being strictly greater than 0 (and the standard deviations being proportional to the number of representing units *a priori* for each class). We converted the priors into energy functionals via Theorem 1:

$$E_{faces}(\sigma_f, m) = \frac{1}{2\sigma_f^2} \sum_{i=0}^{m} i^2$$

which can be simplified by solving the summation:

$$E_{faces}(\sigma_f, m) = \frac{1}{2\sigma_f^2} \frac{m(m+1)(2m+1)}{6}$$

$$E_{houses}(\sigma_h, m) = \frac{1}{2\sigma_h^2} \sum_{i=0}^{m} (i - \mu_h)^2$$

which can be simplified by solving the summation:

$$E_{houses}(\sigma_h, m) = \frac{1}{2\sigma_h^2} \left[\frac{m(m+1)(2m+1)}{6} - \frac{\mu_h}{2} m(m+1) + \mu_h^2 m \right]$$

Finally, we scaled the derivatives according to the size and type of stimuli. This means that topographic bias isn't the only factor which acts on the error derivatives. Thus the topographic bias energy functional we derived before is not enough. However, from the derivation of that energy functional, it is easy to see that the energy functional E_τ is now

$$E_\tau(\sigma_\tau, m, n, \mathbf{w}, f_s, h_s, f, h) = \sum_i^m \sum_j^n \tau(i,j)\psi(f_s, h_s, f, h)E_D(\mathbf{w})$$

where $\psi(s, f, h) = (1-s)^f(s)^h$, with s being the scale in percent, and f and h being indicator variables signifying whether the image is a face or a house. In a simplified form, this becomes:

$$E_\tau(\sigma_\tau, m, n, \mathbf{w}, s, f, h) = \frac{s^h(1-s)^f E_D(\mathbf{w})}{e^{\sigma_\tau^2 - 2}} \sum_i^m \sum_j^n e^{i^2+ij+j^2}$$

The total energy functional is then:

$$E_{eb} = E_\tau(\sigma_\tau, m, n, \mathbf{w}, s, f, h) + E_W(\mathbf{w}, \alpha, m, n) + E_{faces}(\sigma_f, m) + E_{houses}(\sigma_h, m)$$

Note that this is an improper energy functional, because when we take the partial derivative with respect to $\mathbf{w_{ij}}$, we have

$$\frac{\partial E_{eb}}{\partial \mathbf{w_{ij}}} = \frac{\partial E_\tau}{\partial \mathbf{w_{ij}}} + \frac{\partial E_W(\mathbf{w})}{\partial \mathbf{w_{ij}}}$$

$$\frac{\partial E_{eb}}{\partial \mathbf{w_{ij}}} = \psi(f_s, h_s, f, h)\tau(i,j)\frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w_{ij}}} + \alpha \mathbf{w_{ij}}$$

as is apparent, the energy functionals associated with the priors dissapear when we optimize with respect to the weights. However, if we were to optimize with respect to the network architecture, specifically to the number of rows m in the matrix $\mathbf{w}$, these terms would suddenly become very important, and the weight decay term dissapears instead. The energy functional is also composite, so there is a tradeoff between modularity/topographic

bias, variance/generalization, and how much of $\mathbf{w}$ is devoted differentially to the representation of faces and houses. An interesting tradeoff to analyze is that between the units devoted to representing faces and the units representing houses. If $\sigma_f > \sigma_h$, then surprisingly $E_{faces} < E_{houses}$, for all m, which is equivalent to saying that adding more units to represent houses is more costly than for adding face-representing units (since m is the total number of rows, and increasing m increases both the number of units representing faces and the number representing houses). This will become more important in our analysis of the performance of the network, as we actually vary the architecture's hidden layer size. Also note what was said about the topographic bias: the network architecture should be learned optimally *a priori*. This can be done by optimizing with respect to m. Another very interesting tradeoff is between modularity and generalization. If the modularity is very high, then this actually prevents generalization of the network's performance to new faces and houses. Again, the formal properties and computational assumptions of this ideal observer will be analyzed with more rigor in tandem with the results.

2.14 Experimental Design

In our preliminary remarks on the learning behavior of the network, it was mentioned that the number of units in the hidden layer should be as low as possible. Also, the number of units devoted to houses must be as small as possible, with the opposite being true for faces. With this in mind, the number of units in the hidden layer were varied. The ideal observer's hidden layer first was made quite large, $32x16$, and progressively corrupted by reducing the hidden layer size. Recall that the horizontal distance was the only distance important in topographic bias. Therefore reducing the number of rows in the hidden layer while keeping the number of columns constant makes the effect of topographic bias the same on each corrupted ideal observer. This means that we can claim that any changes in network performance are not due to effects from the topographic bias. Any changes in network performance must then be due to a change in the number of units in the hidden

layer. In the language of our energy functional, we varied m and controlled for all other parameters. The hypothesis from our energy functional is that a small hidden layer, with a larger proportion of units selective to faces than to houses, will best match the behavior seen in the activation studies mentioned before.

The ideal observer and the corrupted observers were trained to classify the faces and houses within-category and between-categories with 100% accuracy. In order to study the selectivity and localization of the representation formed by each network, we lesioned a group of three columns of the hidden layer with varying eccentricity (horizontal position in the hidden layer) and recorded the classification accuracy for both faces and houses of the network on novel faces and houses (the test set). Note that the both the training set and the testing set were taken from the same pool of images, and randomly assigned to either set.

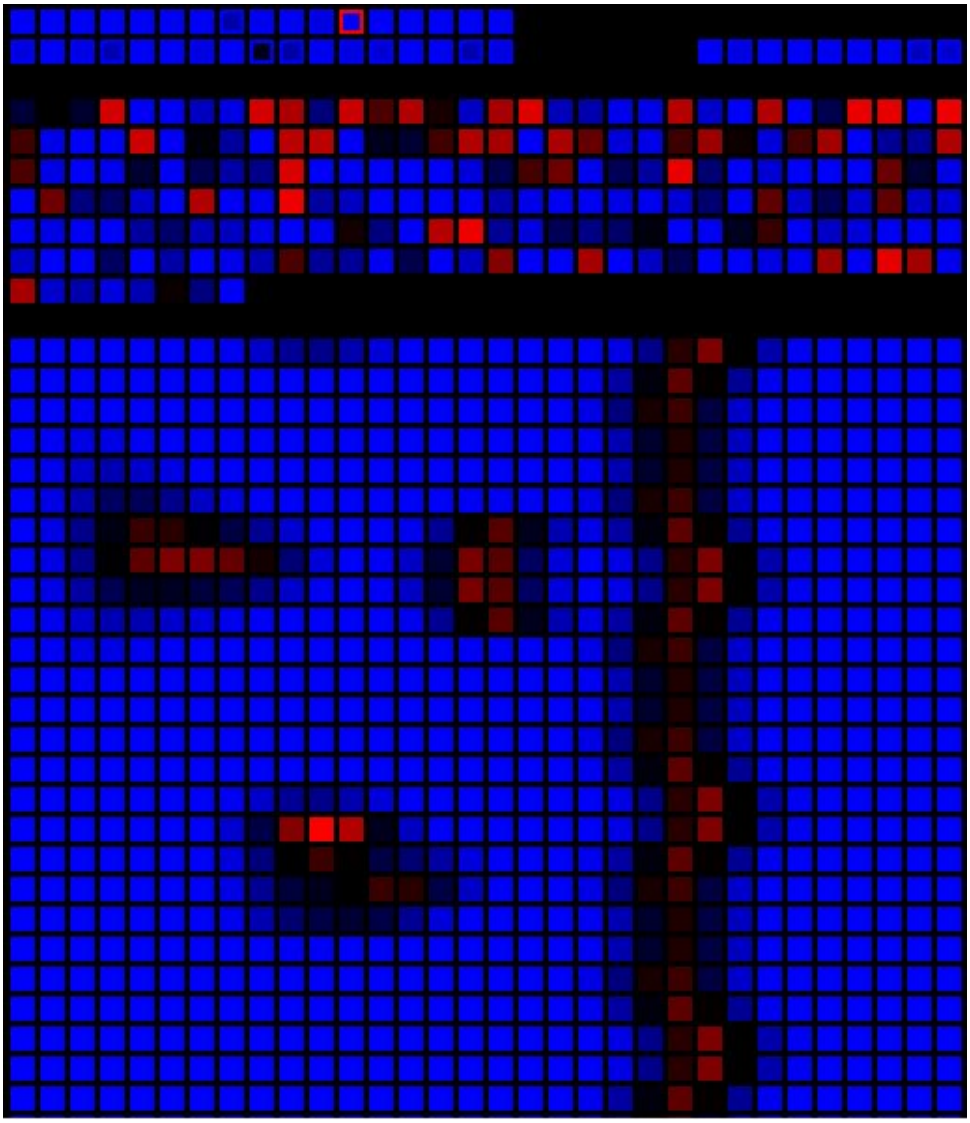

Figure 2.5: This is the network architecture for the ideal observer. The layer of units at the top is the output layer, the group of units under that is the hidden layer, and the units under the hidden layer are the input layer. The absolute value of a unit's activation is the unit's brightness, with red indicating positive activation and blue indicating negative activation. The input layer activations correspond to pixels on a log-polar transformed bitmap of a face. Each column in the input layer represents a different eccentricity, and each row a different orientation (starting with zero degrees at the upper left corner and zero eccentricity at the lower left). Thus the leftmost section of the input layer corresponds to foveal input, and the rightmost section corresponds to peripheral input.

Chapter 3

Results and Analysis

3.1 Results

After constructing the ideal observer and reducing the number of rows in the hidden layer, we tested how the classification error for both faces and houses was localized in eccentricity. The results are shown in Figures 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6. The first four graphs are of the classification error as a function of the eccentricity of the column lesioned for a set number of rows in the hidden layer. The last two graphs are of classification error of all four networks for faces and houses respectively as a function of eccentricity of the lesioned column.

It can be seen that in Figures 3.5 and 3.6 the corrupted ideal observer with $m = 4$ rows in its hidden layer also had the most eccentricity-biased selectivity to faces and houses. That is, the maximal selectivity as a function of eccentricity was attained when the number of rows was minimal. Interestingly, there is an apparent relationship between the eccentricity of the lesioned columns and the classification accuracy. The face-selective area is more central (low in eccentricity), whereas the house-selective area is more peripheral. This is exactly the type of representation shown to exist in human visual cortex by the activation studies. The eccentricity bias does not segregate object representations completely. In fact, it is a topographic, overlapping representation. This pattern also emerges from our ideal observers, as for intermediate eccentricity lesions the impact to classification accuracy is of similar magnitude. This of course also agrees with the eccentricity bias hypothesis.

Finally, recall that one aspect of the energy functional introduced by the architecture is that the number of units used to represent faces is much higher than the number for houses, and this disparity increases with decreasing m. Because the double-dissociation was most apparent in the corrupted observer with the least rows, this implies that the number of units used in face recognition is much greater than the number used for house recognition. This relationship can be seen in Figure 3.4, in which the classification accuracy for faces is

lower than for houses when lesions are done at 68% of the eccentricities. This suggests that face-selective units outnumber house-selective units by almost a ratio of $7:3$. When we consider that each unit in an ideal observer's hidden layer represents tens of thousands of cells, this makes the number of face-selective neurons predicted by the model much greater than the number of house-selective neurons. This in turn implies that the dimensionality of the observer's representation of faces is much higher, meaning that higher-resolution information is used for face recognition than for house recognition. This again agrees with the eccentricity bias hypothesis.

Interestingly, the number of units activated in the input for face images that were learned by the network was relatively low compared to the high representational power of the approximately 44 units (68 percent of 64) which are most face-selective. This implies that there are relatively few units used to represent any one face. In contrast, for house recognition, the number of units active in the input layer for each image of a house was much higher relative to the reduced representational power of the 20 units that are more house-selective. This means that most of the house-selective units are used to represent any given house. What does this mean? Because of the high dimensionality of the representation for faces, a face is easier for us to identify than a specific house. As a consequence, objects that require high-resolution information also are represented very efficiently. Not surprisingly, this has also been found in the outstanding performance of independent components analysis as a classifier for faces (Bartlett and Sejnowski, 2005). An intriguing consequence of this ideal model's behavior is then that we can do face recognition in a way that is optimal in terms of coding efficiency. So, using the language of information theory, perhaps the way objects are represented in the visual system is a function of how efficient of a code we can make of them. There are of course many examples of optimal encodings throughout the visual system, but mostly in early vision (Olshausen and Field, 2002) (Doi et al., 2003) (Laughlin, 1989).

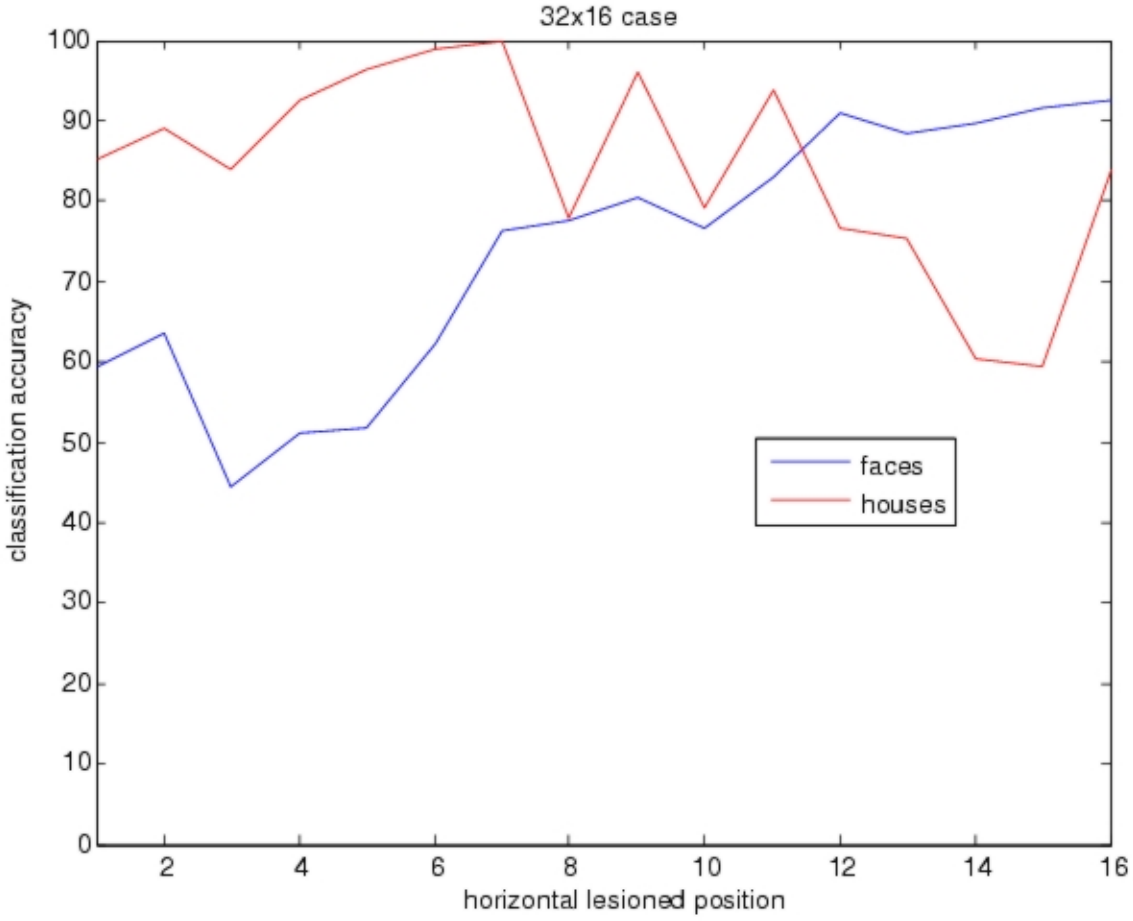

Figure 3.1: This is the classification on novel houses and faces error for the largest (uncorrupted) ideal observer as a function of the eccentricity of the cluster of columns lesioned.

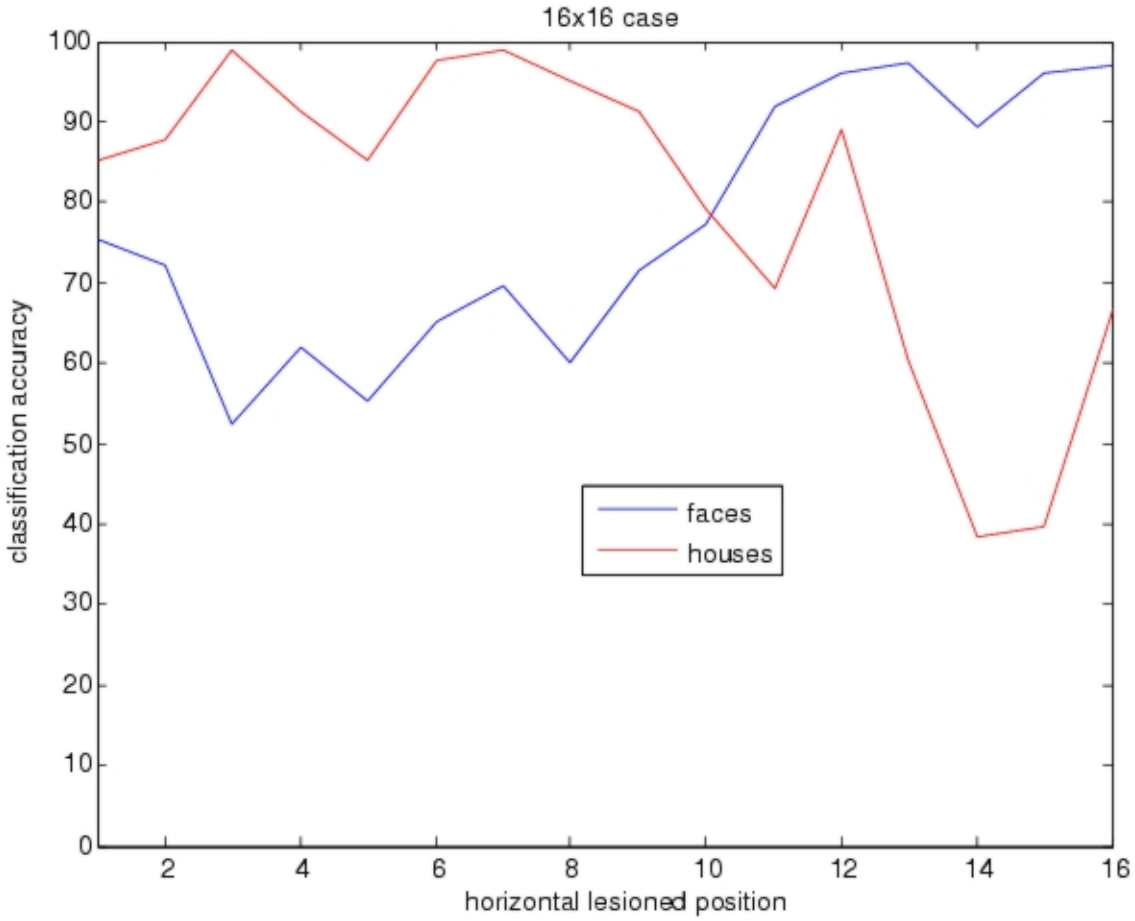

Figure 3.2: This is the classification error on novel houses and faces for the corrupted observer with 16 rows in its hidden layer as a function of the eccentricity of the cluster of columns lesioned.

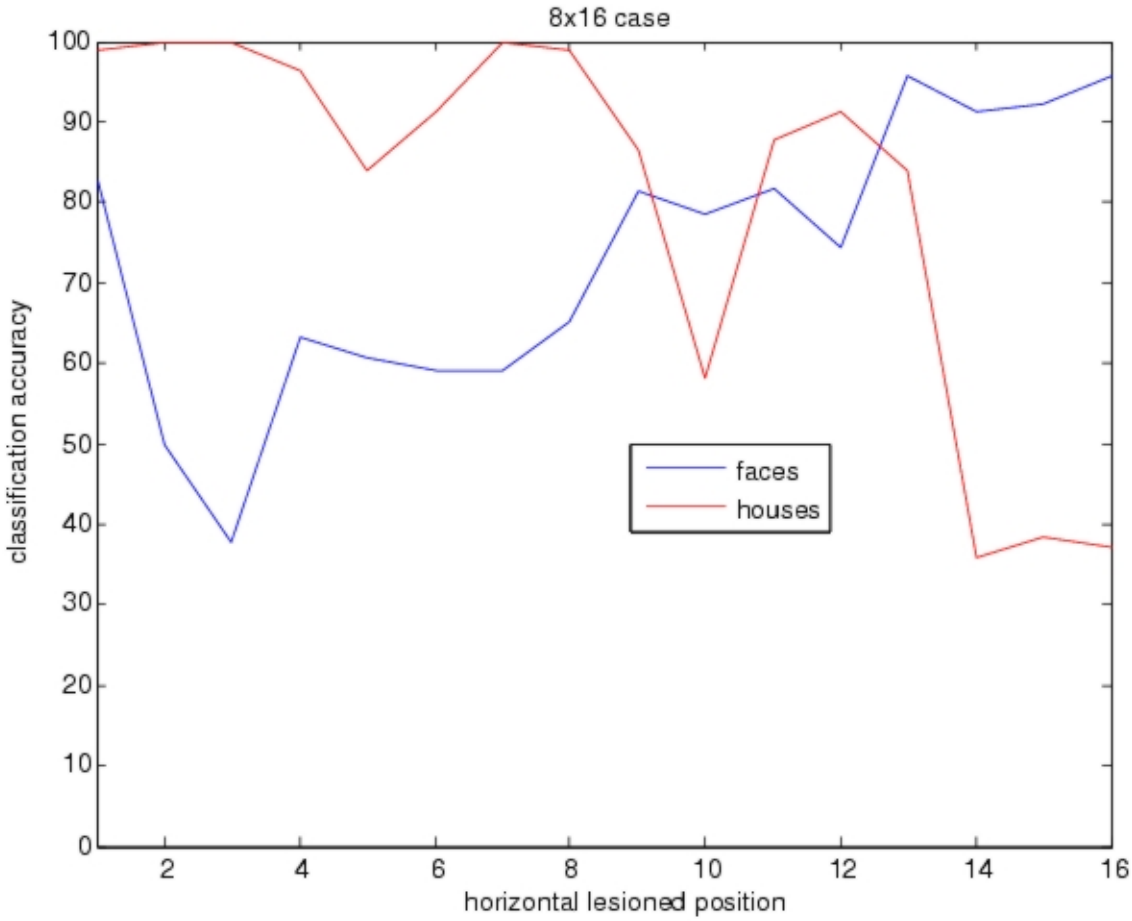

Figure 3.3: This is the classification error on novel houses and faces for the corrupted observer with 8 rows in its hidden layer as a function of the eccentricity of the cluster of columns lesioned.

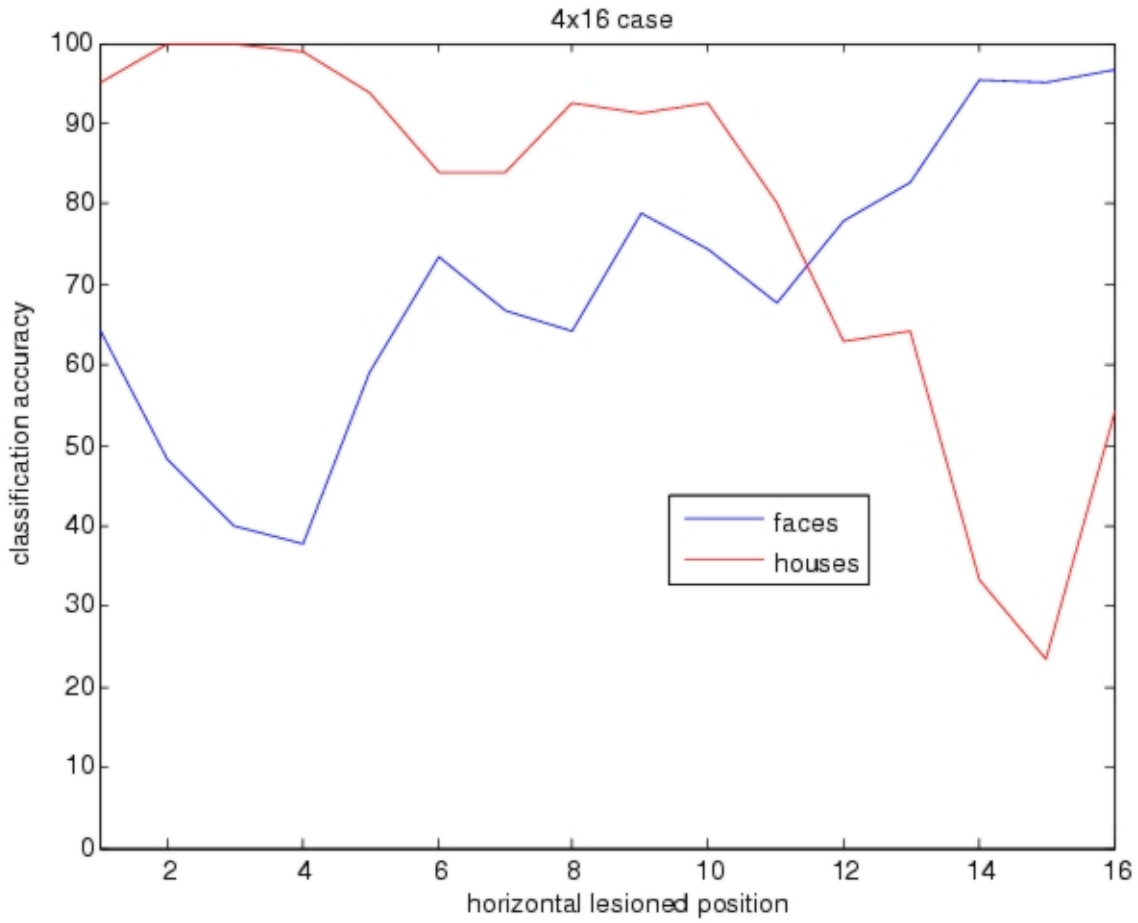

Figure 3.4: This is the classification error on novel houses and faces for the corrupted observer with 4 rows in its hidden layer as a function of the eccentricity of the cluster of columns lesioned.

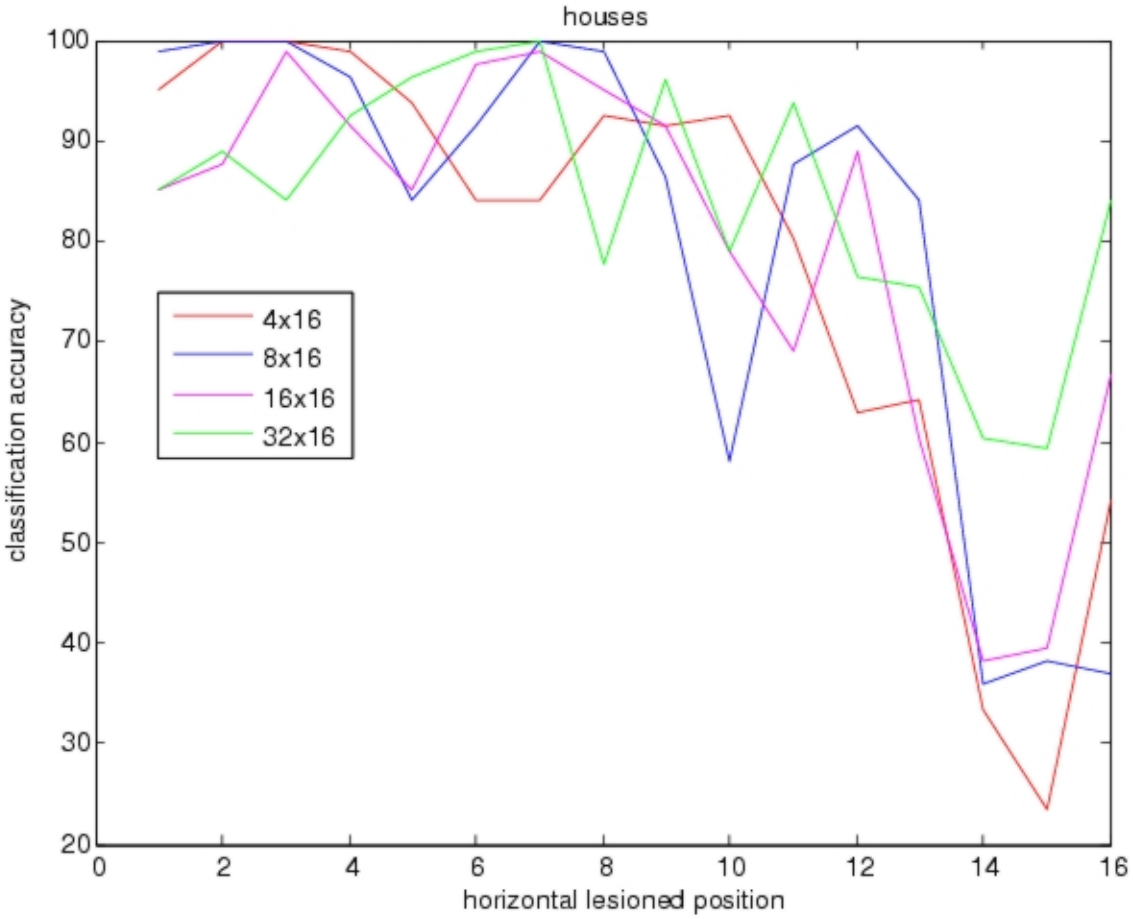

Figure 3.5: This is the classification error for novel houses of all four ideal observers as a function of the eccentricity of the cluster of columns lesioned.

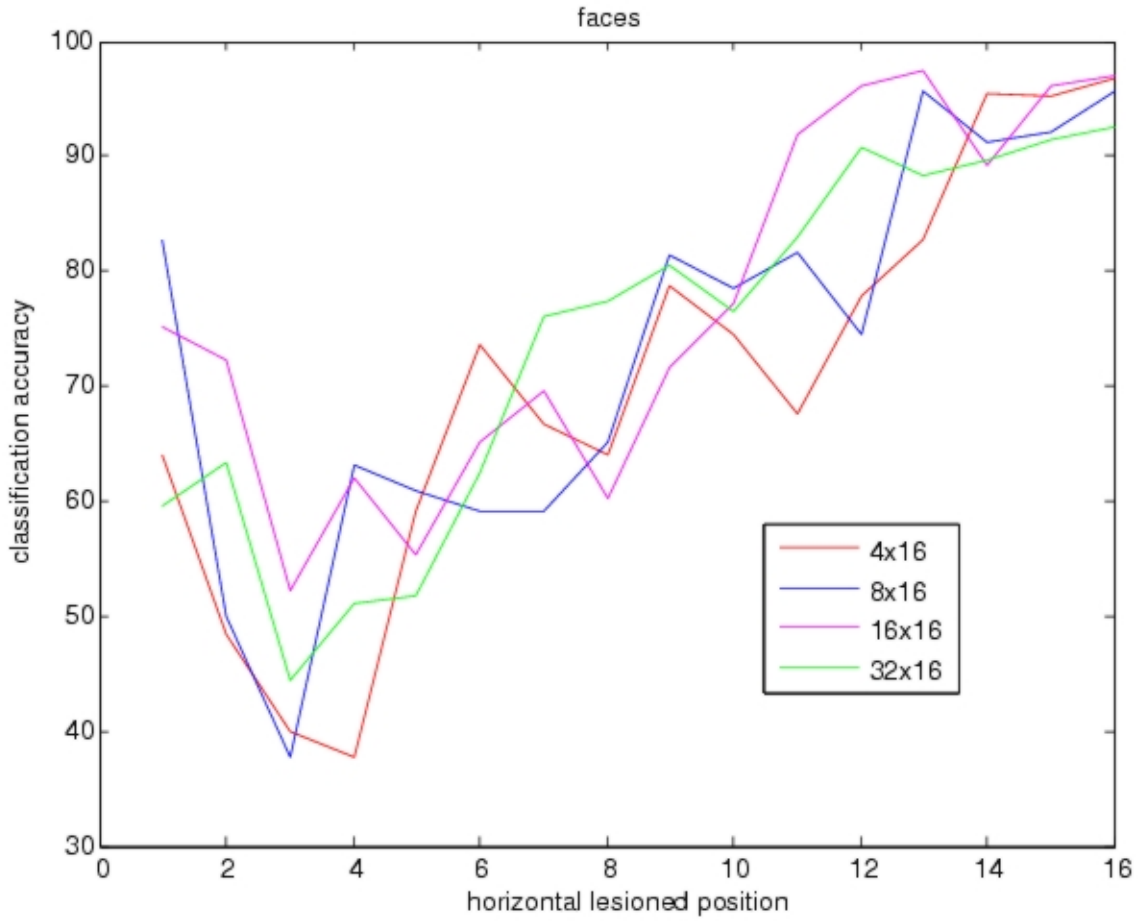

Figure 3.6: This is the classification error for novel faces the largest (uncorrupted) ideal observer as a function of the eccentricity of the cluster of columns lesioned.

3.2 Analysis

Now that we have established that the ideal observer we constructed was indeed in agreement with the activation studies and eccentricity bias hypothesis, we can look at further assumptions that are made in the model which contribute causally to the learning behavior of the network. The most rigorous models of causality are, conveniently, graphical models (Spirtes et al., 2000); (Pearl, 2000); (Druzdzel and Simon, 1993). So if we want a causal model of the learning done by our ideal observer, it suffices to consider its graphical model. Using the hierarchical bayesian methods and considering the value of the energy functional a random variable, we can draw the graphical model pictured in Figure 3.7. The dependencies in this graphical model are just the parameters used in the energy functional which are or could be learned. For example, E_{eb} is the sum of four different energy functionals, which if the energy functionals do not share parameters would make them independent of each other. However, clearly many different energy functionals rely on the number of rows m and the weights w. The graphical model for the learning behavior of the network is quite complicated, so analyzing learning behavior in an ad-hoc way is frought with peril.

First, we observe that graphical models encode conditional dependencies between random variables. This means that we can actually discover new conditional dependencies that were not even considered by ad-hoc approaches. Our great ally in causal analysis is d-separation, which was described before and first derived in the context of causality (Pearl, 2000). Even more significant is that d-separation can be checked algorithmically in a graphical model, almost directly from the definition of an active trail (though see (Shachter, 1998) for a popular alternative).

This means we can easily and methodically discover important causal dependencies in any graphical model. Before summarizing the results from d-separation, we first mention the

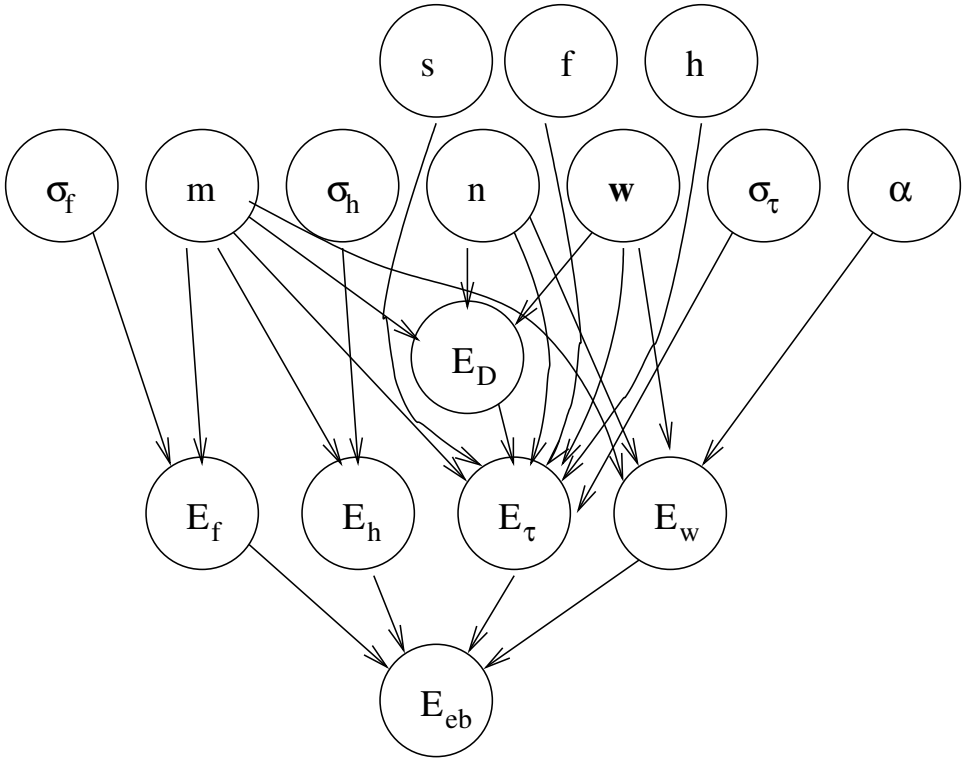

Figure 3.7: Graphical model of the learning that is done by the ideal observer. σ_f is proportional to the number of units selective to faces, σ_h to the number of units selective to houses, m to the number of rows, n to the number of columns, σ_τ to the amount of modularity or locality in the representation, $\mathbf{w}$ to the learned mapping between inputs and responses, s to the scale of the image, f and h indicating whether the image is a face or a house, E_D to the error of the observer's output, and E_w to how narrow-minded or context-specific the observer's learning is.

more obvious causal dependencies. Also, what does observation mean in the context of our ideal observer? When a variable in the graphical model is observed, it means we know what the current value of the variable is. If our analysis of the ideal observer's learning behavior is to be as general as possible, we should consider that we can only know what the value of a variable is when we are not learning with respect to it. That is, if we manipulate a variable directly in the learning process, then it is considered to be unobserved. This is intuitively because if we knew what the value of the variable was, then we wouldn't be attempting to learning it. First, we notice that there are explaining away effects if we observe only E_{eb}, as the network we implement does. That is, knowing with any certainty the modularity, generalization, resources devoted to face recognition, or resources devoted to house recognition reduces other three. Therefore, all four of those energy functionals are dependent in the ideal observer. This argues against hap-hazard optimization of one energy functional without taking the others into account. It could be argued though that this could be an artefact of our model that wouldn't be true of the visual system. To counter that argument, we invoke the powerful d-separation methods described before.

The great power of causal analysis in graphical models is that we can find dependencies that are more general than just the ones seen in our specific ideal observer. So we can even see which causal dependencies exist independent of our specific implementation of the ideal observer. So first we relax the assumption that we know E_{eb}. Now we can discuss dependencies not in our implementation of the ideal model, but are still essential to learning which was shown to be essential to the eccentricity bias hypothesis.

Applying d-separation on Figure 3.7, we find that $\sigma_\tau \perp E_D$. This states that the amount of locality or modularity is always independent of how high the classification error is. This is very good news indeed for the modularity hypothesis.

3.2.1 Known network architecture

If we observe m, then by their d-separation $E_f \perp E_h \perp E_\tau \perp E_D \perp E_W$. Note that we observe m when the network architecture is already determined. That is, when the number of neurons in the brain is set. Therefore, if the number of neurons is constant, modularity is independent of everything else, and should be optimized as Kanwisher prescribes (as much as possible).

Also, if the network architecture is set, then the number of neurons selective to faces and houses has no impact on how local the representation is or how high the classification error is. This suggests that the reason for the high dimensionality of the face representation is not to reduce classification error or to increase modularity, but due to the statistics of face images.

Finally, if m is observed, then s is d-separated from E_f and E_h, so therefore $E_f \perp s|m$ and $E_h \perp s|m$. This is interesting because it suggests that the number of units selective to faces and houses are independent of the scale of the image in our ideal observer. This means that our ideal observer can do scale-invariant face and house recognition.

3.2.2 Known weight matrix

What causal relationships hold if we consider a variant of the ideal observer that has innate knowledge of how to do face and house recognition? In this case w is observed, which under d-separation means $E_\tau \perp E_D \perp E_w$. Therefore, no matter whether m is observed, if w isn't observed, a change in any variable will affect the generalization of what is learned. This is because any two nodes which are not d-separated are causally dependent. the ideal observer's learning is. So recklessly declaring that there is a domain-specific face module is equivalent to saying that domain-specificity and localization are optimized independently from any other factors. This means that if there is no learning in the human brain's object and face recognition processes, then Kanwisher might actually be right about

domain-specific modules for face recognition. There is then an empirically-testable prediction. If learning and synaptic plasticity are demonstrated in the FFA, then our criticism of the modular hypothesis still holds. However, if there is no synaptic plasticity in the FFA, then face recognition is fully modular, domain-specific, and innate. However, the evidence supporting the expertise hypothesis suggests that there is plasticity in this region, and so it is more likely that face recognition is not modular.

3.3 Discussion

It is evident that simplistic hypotheses that are not quantified mathematical models can have major pitfalls. The modularity and expertise hypotheses both are very simple models of object recognition, and thus fail to capture all of the essential features of how the human brain recognizes faces. Eccentricity bias theory provides a specific and mathematical account of face recognition, and our constructed ideal observer model seems to capture the richness of behavior reported in the activation studies. In addition, it can be interpreted that eccentricity bias is also a theory of optimal coding and image statistics. The ideal observer constructed teaches us two things: first, face recognition requires higher dimensional representation and efficient coding; second, object recognition requires a balance of modularity, generalization, units selective to faces, and units selective to houses. Any more naive account will perhaps fail to capture the full range of human object representation.

Graphical models proved to be quite useful for discovering causal relationships in the ideal observer. Even more fascinating than this is the fact that they can be used to turn the ideal observer we constructed into a face and object recognition algorithm which is guaranteed to have the same learning behavior as gave rise to the effects seen in the activation studies. That is, we can "compile" our ideal observer into a face and object recognition algorithm that behaves the same way as the human subjects in the activation

experiments.

Chapter 4

ECCO: Eccentricity-biased Object and Face recognition

4.1 Graphical Optimization Models

There are many pitfalls associated with optimizing energy and loss functionals, and the different methods fail to be sufficiently general or fast to be applied to many different learning problems. Simulated annealing has guaranteed convergence to a global maximum or minimum (Kirkpatrick et al., 1983), but is notorious for being extremely slow and hard to work with, due to the careful tuning that is required for the temperature schedule. There are numerous problems with gradient methods (Steihaug, 1983), as they are notorious for getting stuck in local maxima or minima, and also have free parameters to be tuned. Expectation-maximization, or EM (Dempster et al., 1977) is known to be much slower than gradient-based methods and also converges to local optima. However, it should be said that EM applies more generally than gradient methods, which generally cannot deal well with inherently probabilistic optimization problems. Finally, the more recent methods of convex optimization (Boyd and Vandenberghe, 2004) show great promise, but one of the most often used convex optimization methods, convex hull, suffers from space complexity issues for intermediate results and poor performance with degeneracies in the energy funcional (Avis and Bremner, 1995).

An alternative method, which is both quite general and able to handle degenerate energy functionals, is to use Theorem 1 to translate energy functionals into graphical models and then to do a MPE query to obtain optimal estimates for the free parameters in the energy functional. This method finds optimal values for each of the free parameters, meaning that it is designed to handle multivariate optimization problems. An MPE query on a graphical model gives the most likely assignment to all free parameters given the data. In fact, as shown in chapter 2, section 8, gradient descent is a special case of a MPE query which assumes a uniform prior on all parameters being optimized (MPE is a special case of MAP). These so-called Graphical Optimization Models do not need such assumptions, since a hierarchical model can be made and prior information about the parameters being optimized

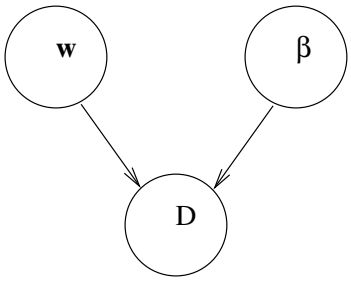

Figure 4.1: This is a simple example of a graphical optimization model. When a MPE query is performed, if the prior probabilities of both $\mathbf{w}$ is uniform, and β is a fixed constant, then this is equivalent to MLE for $\mathbf{w}$, and thus gradient descent.

can be integrated into the optimization model. For example, say we wish to optimize some mapping $\mathbf{w}$ with respect to $E_D(\mathbf{w})$ assuming a learning rate β. Then, according to Theorem 1, the statistical model for this energy functional is:

$$P(D|\mathbf{w},\beta) = \frac{\delta(\mathbf{w})e^{-\beta E_D(\mathbf{w})}}{Z_\beta}$$

which can be translated into a graphical optimization model by translating the only conditional dependency in the model, $D|\mathbf{w},\beta$ (see Figure 4.1) the power of such a way of looking at optimization is that we are not required to assume nothing before we search for optimal parameters. If we have prior knowledge about the infeasibility of a range of values for the parameters, we can reduce the search space substantially. Importantly, we can derive a prior distribution for the parameters based upon the statistics of the data D . A powerful way of doing this for images is by using minimax entropy methods (Zhu et al., 1997); (Huang and Mumford, 1995). Also, because the probability distribution includes the degeneracy function $\delta(\theta)$, we see directly the effect that the degeneracy of the energy functional has on our ability to find optimal estimates. This means we can potentially improve our energy functional and handle degeneracy in a robust way (recall that this was a problem for convex hull algorithms). A future paper will detail further work in this area.

4.1.1 Caveats of graphical optimization models

Of course, graphical optimization models are not a panacea. Finding an exact MPE on arbitrary graphical models is an NP-complete problem (Bylander et al., 1991). However, this means that the general exact inference problem is very hard, not that it is hopeless. In practice there are many approximations which sidestep the computational complexity of MPE. The most often used approximation is loopy belief propagation (Pearl, 1988), which converges to the correct MPE (using the max operation instead of sum) only in the case of an acyclic graph (a polytree, to be precise). The error of the MPE query is actually quite low though for most graphical models, with the error increasing on inclusion of very tight loops and deterministic mappings in the graphical model (Yedidia et al., 2001). In practice if the graph is highly cyclic, double-loop belief propagation can be used (Yuille, 2002). For the sake of sanity though if there are many tight loops in the graphical optimization model corresponding to an energy functional, it is perhaps wiser to either use other approximate algorithms (for example variational methods (Jordan et al., 1999)) or other optimization methods were applicable.

4.2 ECCO: Eccentricity Bias as a Learning algorithm

The desirable properties of the human visual system are well known. The properties of object recognition captured by our ideal observer are scale invariance, optimal classification for different types of objects, and possibly optimal coding properties. All of these properties and more seen to stem from the human representation of objects can be infused in a graphical model. Using Theorem 1, we can construct a graphical optimization model from the energy functional optimized by the ideal observer, and also encode patterns we noticed in what types of assignments of parameters were best.

Recall the definition we used of the energy functional for the ideal observer:

$$E_{eb} = E_\tau(\sigma_\tau, m, n, \mathbf{w}, s, f, h) + E_W(\mathbf{w}, \alpha, m, n) + E_{faces}(\sigma_f, m) + E_{houses}(\sigma_h, m)$$

the statistical model for this energy functional is then, by the chain rule for probability and Theorem 1,

$$P(D|\sigma_\tau, \sigma_f, \sigma_h, f, h, s, \mathbf{w}, A, m, n, \alpha) = P(D|\sigma_\tau, m, n, \mathbf{w}, s, f, h)P(\mathbf{w}|\alpha, A, m, n)P(A|\sigma_f, m)P(A|\sigma_h, m)$$

where A is the total number of rows required to represent the current object. The graphical optimization model given by this factorization of the likelihood is given in Figure 4.2.

To define the probabilities for all terms in the graphical optimization model, we must first recall some definitions.

$$g(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

$$y_h = g(\sum_i x_i w_{ih}) \tag{4.2}$$

$$y_o = \sum_h y_h w_{ho} \tag{4.3}$$

$$y_o = \sum_h g(\sum_i x_i w_{ih}) w_{ho} \tag{4.4}$$

and that we can simplify this with

$$y_o(x_h; \mathbf{w}) = \sum_h g(x_h \mathbf{w}_{ho})$$

where $\mathbf{w}$ is a random matrix. and further recall that the error is defined as

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_h \sum_o (d_o^{(h)} - y_o(x_h; \mathbf{w}))^2$$

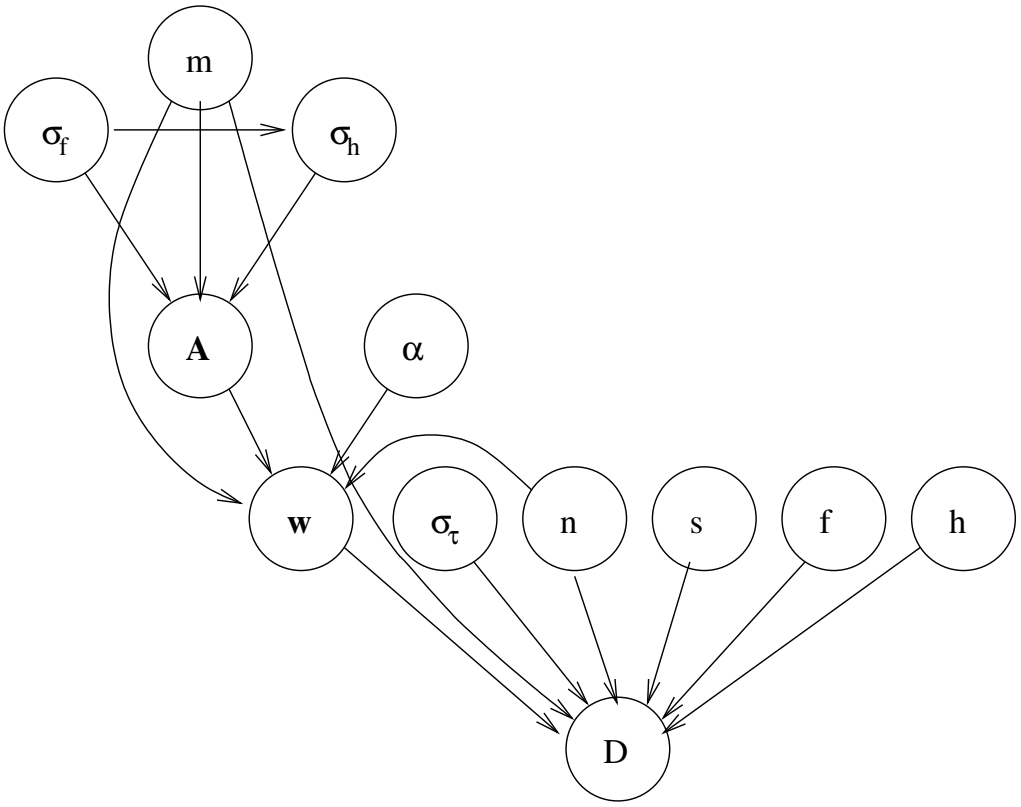

Figure 4.2: Graphical model of the learning that is done by the ideal observer. σ_f is proportional to the number of units selective to faces, σ_h to the number of units selective to houses, m to the number of rows, n to the number of columns, σ_τ to the amount of modularity or locality in the representation, $\mathbf{w}$ to the learned mapping between inputs and responses, s to the scale of the image, f and h indicating whether the image is a face or a house, and A is the number of rows used to represent an object.

with d being a vector provided during training which has a one for the correct identity and a zero for the rest. By Theorem 1,

$$P(D|\sigma_\tau, m, n, \mathbf{w}, s, f, h) = \frac{e^{-\psi(s,f,h)E_D(\mathbf{w})\sum_i^m\sum_j^n\tau(i,j)}}{Z_{E_\tau}}$$

with

$$\tau(i,j) = e^{\frac{-(j-i)^2}{\sigma_\tau^2}}$$

$$\psi(s,f,h) = (1-s)^f(s)^h$$

$$P(\mathbf{w}|\alpha, A, m, n) = \frac{e^{-(1/2)\sum_i^m\sum_j^n\alpha\mathbf{w}_{ij}^2}}{Z_{E_\mathbf{w}}}$$

$$P(A|\sigma_f, m) = \frac{e^{-\frac{1}{2\sigma_f^2}\sum_i^m i^2}}{Z_{E_{faces}}}$$

$$P(A|\sigma_h, m) = \frac{e^{-\frac{1}{2\sigma_h^2}\sum_i^m(i-\mu_h)^2}}{Z_{E_{houses}}}$$

where $\mu_h = 4$ is introduced by the eccentricity bias of houses for peripheral representations. Because this is a bayesian hierarchical model, we can also talk about the distribution of the parameters. By doing this, we can include what we learned from tuning the parameters of the ideal observer. First, recall that the values of m and n we found worked best were 4 and 16 respectively, but that was for a small image, so if the model is to work with larger images, both could be much larger. We can say then that the following holds true of the parameters, with a large uncertainty so that the model can adapt to variable image size:

$m \sim N(4, 100)$

$n \sim N(16, 100)$

The values of σ_f and σ_h were not precisely found, but it was noticed that $3\sigma_f = 7\sigma_h$. Therefore, we actually see that σ_h can be written in terms of σ_f, which is why in Figure 4.2 there is an arrow between the two. We can estimate as we did before that about 44 units are used for faces. This gives the following expressions for the parameters:

$\sigma_f \sim N(44, 1)$

$$P(\sigma_h|\sigma_f) = \frac{3P(\sigma_f)}{7}$$

We can assume that s, f and h are provided by the trainer as well during training. That is, each image is labeled during training with three things, its identity, what type of object it is, and what scale it is. We can calculate each of these probabilities and then use belief propagation to solve the MPE problem for our graphical optimization model. We can use belief propagation because there is only one tight loop in the model and all mappings between random variables are stochastic.

So ECCO, in summary, is a graphical optimization model that gives an estimate of $\mathbf{w}$, m, n, σ_f, and σ_h given training data D that includes images labeled with scale, identity, and category. After training, the learned $\mathbf{w}$ can be used to get the identity L of an image I as follows:

$$L = \mathbf{w}^T I$$

L is a matrix such that $arg\max_{i,j} L(i, j)$ is the the identity of image I.

4.3 Future Work

In future work, optimizing with respect to degeneracy seems like a good direction, and implementing ECCO, possibly applying it to the FERET database of faces which is standard in face recognition. Note that in ECCO we assumed that all energy functionals were non-degenerate, which is not true.

Bibliography

Ackley D, Hinton G, Sejnowski T (1985) A Learning Algorithm for Boltzmann Machines. *Cognitive Science* 9:147–169.

Avis D, Bremner D (1995) How good are convex hull algorithms? *Proceedings of the eleventh annual symposium on Computational geometry* pp. 20–28.

Bartlett M, Sejnowski T (2005) *Independent components of face images: A representation for face recognition*.

Bentin S (2000) STRUCTURAL ENCODING AND IDENTIFICATION IN FACE PROCESSING: ERP EVIDENCE FOR SEPARATE MECHANISMS. *Cognitive Neuropsychology* 17:35–55.

Berger J (1993) *Statistical Decision Theory and Bayesian Analysis* Springer.

Bishop C (1996) *Neural Networks for Pattern Recognition* Oxford University Press, USA.

Boyd S, Vandenberghe L (2004) *Convex Optimization* Cambridge University Press.

Bylander T, Allemang D, Tanner M, Josephson J (1991) The Computational Complexity of Abduction. *Artificial Intelligence* 49:25–60.

Carey S, Diamond R (1977) From piecemeal to configurational representation of faces. *Science* 195:312–4.

Clark V, Keil K, Maisog J, Courtney S, Ungerleider L, Haxby J (1996) Functional Magnetic Resonance Imaging of Human Visual Cortex during Face Matching: A Comparison with Positron Emission Tomography. *NeuroImage* 4:1–15.

Courtney S, Ungerleider L, Keil K, Haxby J (1997) Transient and sustained activity in a distributed neural system for human working memory. *Nature* 386:608–611.

Damasio A (1990) Category-related recognition defects as a clue to the neural substrates of knowledge. *Trends Neurosci* 13:95–8.

Damasio A, Damasio H, Van Hoesen G (1982) Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology* 32:331–41.

Damasio A, Tranel D, Damasio H (1990) Face agnosia and the neural substrates of memory. *Annu Rev Neurosci* 13:89–109.

De Renzi E (1986) Current issues in prosopagnosia. *Aspects of face processing* 28:243–252.

Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1–38.

DeYoe E, Carman G, Bandettini P, Glickman S, Wieser J, Cox R, Miller D, Neitz J (1996) Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences* 93:2382–2386.

Diamond R, Carey S (1986) Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General* 115:107–117.

Doi E, Inui T, Lee T, Wachtler T, Sejnowski T (2003) Spatiochromatic Receptive Field Properties Derived from Information-Theoretic Analyses of Cone Mosaic Responses to Natural Scenes. *Neural Computation* 15:397–417.

Druzdzel M, Simon H (1993) Causality in Bayesian belief networks. *Uncertainty in Artificial Intelligence* 9:3–11.

Duda R, Hart P, Stork D (2001) *Pattern classification* Wiley New York.

Duncan R, Boynton G (2003) Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron* 38:659–671.

Eimer M (1998) Does the face-specific N170 component reflect the activity of a specialized eye detector. *NeuroReport* 9:2945–2948.

Eimer M (2000) Effects of face inversion on the structural encoding and recognition of faces: Evidence from event-related brain potentials. *Cognitive Brain Research* 10:145–158.

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. *Nature* 392:598–601.

Farah M (2004) *Visual agnosia* MIT Press Cambridge, Mass.

Farah M, Levinson K, Klein K (1995) Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia* 33:661–74.

Farah M, Wilson K, Drain M, Tanaka J (1998) What is" special" about face perception? *Psychol Rev* 105:482–98.

Fodor J (1983) *Modularity of Mind* Bradford Book.

Gauthier I, Anderson A, Tarr M, Skudlarski P, Gore J (1997) Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Curr Biol* 7:645–51.

Gauthier I, Skudlarski P, Gore J, Anderson A (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience* 3:191–197.

Gauthier I, Tarr M (1997) Becoming a Greeble expert: Exploring mechanisms for face recognition. *Vision Research* 37:1673–1682.

Gauthier I, Tarr M (2002) Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance* 28:431–446.

Gauthier I, Tarr M, Anderson A, Skudlarski P, Gore J (1999) Activation of the middle fusiform'face area'increases with expertise in recognizing novel objects. *Nature Neuroscience* 2:568–573.

Geisler W (2003) Ideal observer analysis. *The visual neurosciences* pp. 825–837.

Good I (1965) *The estimation of probabilities: An essay on modern Bayesian methods* The MIT Press.

Good I (1987) Hierarchical Bayesian and empirical Bayesian methods (letter). *American Statistician* 41:92.

Grill-Spector K, Kushnir T, Edelman S, Itzchak Y, Malach R (1998a) Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron* 21:191±202.

Grill-Spector K, Kushnir T, Hendler T, Edelman S, Itzchak Y, Malach R (1998b) A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human Brain Mapping* 6:316–328.

Grill-Spector K, Malach R (2004) THE HUMAN VISUAL CORTEX. *Annual Review of Neuroscience* 27:649–677.

Haig N (1984) The effect of feature displacement on face recognition. *Perception* 13:505–12.

Hasson U, Harel M, Levy I, Malach R (2003) Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron* 37:1027–1041.

Hasson U, Levy I, Behrmann M, Hendler T, Malach R (2002) Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* 34:479–490.

Haxby J, Horwitz B, Ungerleider L, Maisog J, Pietrini P, Grady C (1994) The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience* 14:6336–6353.

Haxby J et al. (2001) Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* 293:2425–2430.

Haxby J, Ungerleider L, Clark V, Schouten J, Hoffman E, Martin A (1999) The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22:189–199.

Henke K, Schweinberger S, Grigo A, Klos T, Sommer W (1998) Specificity of face recognition: recognition of exemplars of non-face objects in prosopagnosia. *Cortex* 34:289–96.

Hinton G, Shallice T (1991) Lesioning an attractor network: investigations of acquired dyslexia. *Psychol Rev* 98:74–95.

Hopfield J (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences* 79:2554–2558.

Hosie J, Ellis H, Haig N (1988) The effect of feature displacement on the perception of well-known faces. *Perception* 17:461–74.

Huang J, Mumford D (1995) Statistics of Natural Images and Models. *log (Histogram)* 7:6.

Ishai A, Ungerleider L, Martin A, Schouten J, Haxby J (1999) Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences* 96:9379–9384.

Jacobs R, Jordan M (1992) Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience* 4:323–336.

Jordan M (1999) *Learning in graphical models* MIT Press Cambridge, MA, USA.

Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37:183–233.

Kanwisher N (2000) Domain specificity in face perception. *Nature Neuroscience* 3:759–763.

Kanwisher N, McDermott J, Chun M (1997) The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience* .

Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. *Science* 220:671–680.

Knill D, Richards W (1996) *Perception as Bayesian Inference* Cambridge University Press.

Laughlin S (1989) The role of sensory adaptation in the retina. *Journal of Experimental Biology* 146:39–62.

Laughlin S, Sejnowski T (2003) Communication in Neuronal Networks. *Science* 301:1870.

Levy I, Hasson U, Avidan G, Hendler T, Malach R (2001) Center-periphery organization of human object areas. *Nature Neuroscience* 4:533–539.

Levy I, Hasson U, Harel M, Malach R (2004) Functional analysis of the periphery effect in human building related areas. *Human Brain Mapping* 22:15–26.

MacKay D (1992) A practical Bayesian framework for backpropagation networks. *Neural Computation* 4:448–472.

Makela P, Nasanen R, Rovamo J, Melmoth D (2001) Identification of facial images in peripheral vision. *Vision Res* 41:599–610.

Malach R, Levy I, Hasson U (2002) The topography of high-order human object areas. *Psychologische Beitrage* 42:201–212.

McCarthy G, Puce A, Gore J, Allison T (1997) Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience* 9:604–609.

McNeil J, Warrington E (1992) *Prosopagnosia: a face-specific disorder.*

Moscovitch M, Winocur G, Behrmann M (1997) What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience* 9:555–604.

Mumford D (1994) The Bayesian rationale for energy functionals. *Geometry-Driven Diffusion in Computer Vision* pp. 141–153.

Nelson M, Bower J (1990) Brain maps and parallel computers. *Trends Neurosci* 13:403–8.

Olshausen B, Field D (2002) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–609.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann.

Pearl J (2000) *Causality: Models, Reasoning, and Inference* Cambridge University Press.

Pietrini P, Furey M, Ricciardi E, Gobbini M, Wu W, Cohen L, Guazzelli M, Haxby J (2004) Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences* 101:5658–5663.

Plaut D (1995) Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology* 17:291–321.

Plaut D, Shallice T (1993) Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* 10:377–500.

Puce A, Allison T, Asgari M, Gore J, McCarthy G (1996) Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of Neuroscience* 16:5205±5215.

Puce A, Allison T, Gore J, McCarthy G (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology* 74:1192–1199.

Rossion B, Gauthier I, Tarr M, Despland P, Bruyer R, Linotte S, Crommelinck M (2000) The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: An electrophysiological account of face-specific processes in the human brain. *NeuroReport* 11:69–74.

Schrodinger E (1989) *Statistical Thermodynamics* Dover Publications.

Sereno M, Dale A, Reppas J, Kwong K, Belliveau J, Brady T, Rosen B, Tootell R (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268:889–93.

Sergent J, Ohta S, MacDonald B et al. (1992) Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain* 115:15–36.

Shachter R (1998) Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)* pp. 480–487.

Simoncelli E, Olshausen B (2001) Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience* 24:1193–1216.

Spirtes P, Glymour C, Scheines R (2000) *Causation, prediction, and search* MIT Press Cambridge, Mass.

Steihaug T (1983) The Conjugate Gradient Method and Trust Regions in Large Scale Optimization. *SIAM Journal on Numerical Analysis* 20:626–637.

Tanaka J, Farah M (1997) *Parts and wholes in face recognition.*

Tanaka J, Gauthier I (1997) Expertise in object and face recognition. *Psychology of learning and motivation* 36:83–125.

Tarr M, Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* 3:764–769.

Tootell R, Mendola J, Hadjikhani N, Ledden P, Liu A, Reppas J, Sereno M, Dale A (1985) Functional Analysis of V3A and Related Areas in Human Visual Cortex .

Tsao D, Freiwald W, Knutsen T, Mandeville J, Tootell R (2003) Faces and objects in macaque cerebral cortex. *Nature Neuroscience* 6:989–995.

Valentine T (1989) *Upside-down faces: a review of the effect of inversion upon face recognition.*

Yedidia J, Freeman W, Weiss Y (2001) Generalized belief propagation. *Advances in Neural Information Processing Systems* 13:689–695.

Yuille A (2002) CCCP Algorithms to Minimize the Bethe and Kikuchi Free Energies: Convergent Alternatives to Belief Propagation. *Neural Computation* 14:1691–1722.

Zhu S, Wu Z, Mumford D (1997) Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation* 9:1627–1660.