FACE IMAGE ANALYSIS BY UNSUPERVISED LEARNING

MARIAN STEWART BARTLETT
Institute for Neural Computation
University of California, San Diego

Kluwer Academic Publishers
Boston/Dordrecht/London

Contents

Foreword

Computers are good at many things that we are not good at, like sorting a long list of numbers and calculating the trajectory of a rocket, but they are not at all good at things that we do easily and without much thought, like seeing and hearing. In the early days of computers, it was not obvious that vision was a difficult problem. Today, despite great advances in speed, computers are still limited in what they can pick out from a complex scene and recognize. Some progress has been made, particularly in the area of face processing, which is the subject of this monograph.

Faces are dynamic objects that change shape rapidly, on the time scale of seconds during changes of expression, and more slowly over time as we age. We use faces to identify individuals, and we rely of facial expressions to assess feelings and get feedback on the how well we are communicating. It is disconcerting to talk with someone whose face is a mask. If we want computers to communicate with us, they will have to learn how to make and assess facial expressions. A method for automating the analysis of facial expressions would be useful in many psychological and psychiatric studies as well as have great practical benefit in business and forensics.

The research in this monograph arose through a collaboration with Paul Ekman, which began 10 years ago. Dr. Beatrice Golomb, then a postdoctoral fellow in my laboratory, had developed a neural network called Sexnet, which could distinguish the sex of person from a photograph of their face (Golomb et al., 1991). This is a difficult problem since no single feature can be used to reliably make this judgment, but humans are quite good at it. This project was the starting point for a major research effort, funded by the National Science Foundation, to automate the Facial Action Coding System (FACS), developed by Ekman and Friesen (1978). Joseph Hager made a major contribution in the early stages of this research by obtaining a high quality set of videos of experts who could produce each facial action. Without such a large dataset of labeled

images of each action it would not have been possible to use neural network learning algorithms.

In this monograph, Dr. Marian Stewart Bartlett presents the results of her doctoral research into automating the analysis of facial expressions. When she began her research, one of the methods that she used to study the FACS dataset, a new algorithm for Independent Component Analysis (ICA), had recently been developed, so she was pioneering not only facial analysis of expressions, but also the initial exploration of ICA. Her comparison of ICA with other algorithms on the recognition of facial expressions is perhaps the most thorough analysis we have of the strengths and limits ICA.

Much of human learning is unsupervised; that is, without the benefit of an explicit teacher. The goal of unsupervised learning is to discover the underlying probability distributions of sensory inputs (Hinton and Sejnowski, 1999). Or as Yogi Berra once said, "You can observe a lot just by watchin'." The identification of an object in an image nearly always depends on the physical causes of the image rather than the pixel intensities. Unsupervised learning can be used to solve the difficult problem of extracting the underlying causes, and decisions about responses can be left to a supervised learning algorithm that takes the underlying causes rather than the raw sensory data as its inputs.

Several types of input representation are compared here on the problem of discriminating between facial actions. Perhaps the most intriguing result is that two different input representations, Gabor filters and a version of ICA, both gave excellent results that were roughly comparable with trained humans. The responses of simple cells in the first stage of processing in the visual cortex of primates are similar to those of Gabor filters, which form a roughly statistically independent set of basis vectors over a wide range of natural images (Bell and Sejnowski, 1997). The disadvantage of Gabor filters from an image processing perspective is that they are computationally intensive. The ICA filters, in contrast, are much more computationally efficient, since they were optimized for faces. The disadvantage is that they are too specialized a basis set and could not be used for other problems in visual pattern discrimination.

One of the reasons why facial analysis is such a difficult problem in visual pattern recognition is the great variability in the images of faces. Lighting conditions may vary greatly and the size and orientation of the face make the problem even more challenging. The differences between the same face under these different conditions are much greater than the differences between the faces of different individuals. Dr. Bartlett takes up this challenge in Chapter 7 and shows that learning algorithms may also be used to help overcome some of these difficulties.

The results reported here form the foundation for future studies on face analysis, and the same methodology can be applied toward other problems in visual recognition. Although there may be something special about faces, we

may have learned a more general lesson about the problem of discriminating between similar complex shapes: A few good filters are all you need, but each class of object may need a quite different set for optimal discrimination.

Terrence J. Sejnowski
La Jolla, CA

Chapter 1

SUMMARY

One of the challenges of teaching a computer to recognize faces is that we do not know a priori which features and which high order relations among those features to parameterize. Our insight into our own perceptual processing is limited. For example, image features such as the distance between the eyes or fitting curves to the eyes give only moderate performance for face recognition by computer. Much can be learned about image recognition from biological vision. A source of information that appears to be crucial for shaping biological vision is the statistical dependencies in the visual environment. This information can be extracted through unsupervised learning[1]. Unsupervised learning finds adaptive image features that are specialized for a class of images, such as faces.

This book explores adaptive approaches to face image analysis. It draws upon principles of unsupervised learning and information theory to adapt processing to the immediate task environment. In contrast to more traditional approaches to image analysis in which relevant structure is determined in advance and extracted using hand-engineered techniques, this book explores methods that learn about the image structure directly from the image ensemble and/or have roots in biological vision. Particular attention is paid to unsupervised learning techniques for encoding the statistical dependencies in the image ensemble.

Horace Barlow has argued that redundancy in the sensory input contains structural information about the environment. Completely non-redundant stimuli are indistinguishable from random noise, and the percept of structure is

[1] "Unsupervised" means that there is no explicit teacher. Object labels and correct answers are not provided during learning. Instead, the system learns through a general objective function or set of update rules.

driven by the dependencies (Barlow, 1989). Bars and edges are examples of such regularities in vision. It has been claimed that the goal of both unsupervised learning, and of sensory coding in the neocortex, is to learn about these redundancies (Barlow, 1989; Field, 1994; Barlow, 1994). Learning mechanisms that encode the dependencies that are expected in the input and remove them from the output encode important structure in the sensory environment. Such mechanisms fall under the rubric of redundancy reduction.

Redundancy reduction has been discussed in relation to the visual system at several levels. A first-order redundancy is mean luminance. Adaptation mechanisms take advantage of this nonrandom feature by using it as an expected value, and expressing values relative to it (Barlow, 1989). The variance, a second-order statistic, is the luminance contrast. Contrast appears to be encoded relative to the mean contrast, as evidenced by contrast gain control mechanisms in V1 (Heeger, 1992). Principal component analysis is a way of encoding second order dependencies in the input by rotating the axes to correspond to directions of maximum covariance. Principal component analysis provides a dimensionality-reduced code that separates the correlations in the input. Atick and Redlich (Atick and Redlich, 1992) have argued for such decorrelation mechanisms as a general coding strategy for the visual system.

This book argues that statistical regularities contain important information for high level visual functions such as face recognition. Some of the most successful algorithms for face recognition are based on learning mechanisms that are sensitive to the correlations in the face images. Representations such as "eigenfaces" (Turk and Pentland, 1991) and "holons" (Cottrell and Metcalfe, 1991), are based on principal component analysis (PCA), which encodes the correlational structure of the input, but does not address high-order statistical dependencies. High order dependencies are relationships that cannot be cannot be captured by a linear predictor. A sine wave $y = sin(x)$ is such an example. The correlation between x and y is zero, yet y is clearly dependent on x. In a task such as face recognition, much of the important information may be contained in high-order dependencies. Independent component analysis (ICA) (Comon, 1994) is a generalization of PCA which learns the high-order dependencies in the input in addition to the correlations. An algorithm for separating the independent components of an arbitrary dataset by information maximization was recently developed (Bell and Sejnowski, 1995). This algorithm is an unsupervised learning rule derived from the principle of optimal information transfer between neurons (Laughlin, 1981; Linsker, 1988; Atick and Redlich, 1992). This book applies ICA to face image analysis and compares it to other representations including eigenfaces and Gabor wavelets.

Desirable filters may be those that are adapted to the patterns of interest and capture interesting structure (Lewicki and Sejnowski, 2000). The more the dependencies that are encoded, the more structure that is learned. Information

theory provides a means for capturing interesting structure. Information maximization leads to an efficient code of the environment, resulting in more learned structure. Such mechanisms predict neural codes in both vision (Olshausen and Field, 1996a; Bell and Sejnowski, 1997; Wachtler et al., 2001) and audition (Lewicki and Olshausen, 1999).

Chapter 2 reviews unsupervised learning and information theory, including Hebbian learning, PCA, mimimum entropy coding, and ICA. Relationships of these learning objectives to biological vision are also discussed. Self-organization in visual development appears to be mediated by learning mechanisms sensitive to the dependencies in the input. Chapter 3 develops representations for face recognition based on statistically independent components of face images. The ICA algorithm was applied to a set of face images under two architectures, one which separated a set of independent images across spatial location, and a second which found a factorial feature code across images. Both ICA representations were superior to the PCA representation for recognizing faces across sessions and changes in expression. A combined classifier that took input from both ICA representations outperformed PCA for recognizing images under all conditions tested.

Chapter 4 reviews automated facial expression analysis and introduces the Facial Action Coding System (Ekman and Friesen, 1978). Chapters 5 and 6 compare image representations for facial expression analysis, and demonstrate that learned representations based on redundancy reduction of the graylevel face image ensemble are powerful for face image analysis. Chapter 5 showed that PCA, which encodes second-order dependencies through unsupervised learning, gave better recognition performance than a set of hand-engineered feature measurements. The results also suggest that hand-engineered features plus principal component representations may be superior to either one alone, since their performances may be uncorrelated.

Chapter 6 compared the ICA representation described above to more than eight other image representations for facial expression analysis. These included analysis of facial motion through estimation of optical flow; holistic spatial analysis based on second-order image statistics such as principal component analysis, local feature analysis, and linear discriminant analysis; and representations based on the outputs of local filters, such as a Gabor wavelet representations and local PCA. These representations were implemented and tested by my colleague, Gianluca Donato. Performance of these systems was compared to naive and expert human subjects. Best performance was obtained using the Gabor wavelet representation and the independent component representation, which both achieved 96% accuracy for classifying twelve facial actions. The results provided converging evidence for the importance of possessing local filters, high spatial frequencies, and statistical independence for classifying facial actions. Relationships between Gabor filters and independent

component analysis have been demonstrated (Bell and Sejnowski, 1997; Simoncelli, 1997).

Chapter 7 addresses representations of faces that are invariant to changes such as an alteration in expression or pose. Temporal redundancy contains information for learning invariances[2] . Different views of a face tend to appear in close temporal proximity as the person changes expression, pose, or moves through the environment. There are several synaptic mechanisms that might depend on the correlation between synaptic input at one moment, and post-synaptic depolarization at a later moment. Chapter 7 modeled the development of viewpoint invariant responses to faces from visual experience in a biological system by encoding spatio-temporal dependencies. The simulations combined temporal smoothing of activity signals with Hebbian learning (Földiák, 1991) in a network with both feed-forward connections and a recurrent layer that was a generalization of a Hopfield attractor network. Following training on sequences of graylevel images of faces as they changed pose, multiple views of a given face fell into the same basin of attraction, and the system acquired representations of faces that were approximately viewpoint invariant.

These results support the theory that employing learning mechanisms that encode dependencies in the input and remove them from the output is a good strategy for object recognition. A representation based on the second-order dependencies in the face images outperformed a representation based on a set of hand-engineered feature measurements for facial expression recognition, and a representation that separated the high order dependencies in addition to the second-order dependencies outperformed representations that separated only the second-order dependencies for both identity recognition and expression recognition. In addition, learning strategies that encoded the spatio-temporal redundancies in the input extracted structure relevant to visual invariances.

[2]"Invariance" in vision refers to the consistency of object identity despite alterations in the input due to translation, rotation, changes in lighting, and changes in scale. One goal is to learn object representations that are unaltered by (invariant to) such changes in the input

Chapter 2

INTRODUCTION

1. UNSUPERVISED LEARNING IN OBJECT REPRESENTATIONS

How can a perceptual system learn to recognize properties of its environment without being told which features it should analyze, or whether its decisions are correct? When there is no external teaching signal to be matched, some other goal is required to force a perceptual system to extract underlying structure. Unsupervised learning is related to Gibson's concept of discovering "affordances" in the environment (Gibson, 1986). Structure and information are afforded by the external stimulus, and it is the task of the perceptual system to discover this structure. The perceptual system must learn about the underlying physical causes of observed images. One approach to self-organization is to build generative models that are likely to have produced the observed data. The parameters of these generative models are adjusted to optimize the likelihood of the data within constraints such as basic assumptions about the model architecture. A second class of objectives is related to information preservation and redundancy reduction. These approaches are reviewed here. The two approaches to unsupervised learning are not mutually exclusive, and it is often possible, as will be seen below, to ascribe a generative architecture to an information preservation objective, and to build generative models with objectives of information preservation. See (Becker and Plumbley, 1996) for a thorough discussion of unsupervised learning. Hinton and Sejnowski's *Unsupervised Learning: Foundations of Neural Computation* (Hinton and Sejnowski, 1999) contains an anthology of many of the works reviewed in this chapter. A recommended background text is Dana Ballard's *Introduction to Natural Computation* (Ballard, 1997).

1.1. Generative models

One approach to unsupervised learning attempts to develop a representation of the data by characterizing its underlying probability distribution. In this approach, a prior model Φ, is assumed which constrains the general form of the probability density function. The particular model parameters are then found by maximizing the likelihood of the model having generated the observed data. A mixture of Gaussians model, for example, assumes that each data point was generated by a combination of causes ϕ_i, where each cause has a Gaussian distribution with a mean u_i, variance σ_i, and prior probabilities or mixing proportions, π_i. The task is to learn the parameters (u_i, σ_i, π_i) for all i that were most likely to have generated the observed data.

Let $\mathbf{x} = [x_1 ... x_n]$ denote the observed data where the n samples are independent. The probability of the data given the model is given by

$$P(\mathbf{x}|\Phi) = \sum_i P(\mathbf{x}|\phi_i)P(\phi_i) \quad (2.1)$$

$$= \prod_j \sum_i P(x_j|\phi_i)P(\phi_i) \quad (2.2)$$

The probability of the data is defined in terms of the prior probability of each of the submodels $P(\phi_i)$ and the posterior probability of the data given the submodel, $P(\mathbf{x}|\phi_i)$, where ϕ_i is defined as (u_i, σ_i, π_i). The parameters of each of the submodels, (u_i, σ_i, π_i), are found by performing gradient ascent on 2.2. The log probability, or likelihood, is usually maximized in order to facilitate calculation of the partial derivatives of 2.2 with respect to each of the parameters. Such models fall into the class of "generative" models, in which the model is chosen as the one most likely to have generated the observed data.

Maximum likelihood models are a form of a Bayesian inference model (Knill and Richards, 1996). The probability of the model given the data is given by

$$P(\Phi|\mathbf{x}) = \frac{P(\mathbf{x}|\Phi)P(\Phi)}{P(\mathbf{x})} \quad (2.3)$$

The maximum likelihood cost function maximizes $P(\mathbf{x}|\Phi)$, which, under the assumption of a uniform prior on the model $P(\Phi)$, also maximizes $P(\Phi|\mathbf{x})$, since $P(\mathbf{x})$ is just a scaling factor.

A variant of the mixture of Gaussians generative model is maximum likelihood competitive learning (Nowlan, 1990). As in the mixture of Gaussians model, the posterior probability $p(x_j|\phi_i)$ is given by a Gaussian with center u_i. The prior probabilities of the submodels $P(\phi_i)$, however, are learned from the data as a weighted sum of the input data, passed through a soft-maximum competition. These prior probabilities give the mixing proportions, π_i.

In generative models, the model parameters are treated as network weights in an unsupervised learning framework. There can be relationships between the update rules obtained from the partial derivative of such objective functions and

other unsupervised learning rules, such as Hebbian learning (discussed below in Section 1.6). For example, the update rule for maximum likelihood competitive learning (Nowlan, 1990) consists of a normalized Hebbian component and a weight decay.

A limitation of generative models is that for all but the simplest models, each pattern can be generated in exponentially many ways and it becomes intractable to adjust the parameters to maximize the probability of the observed patterns. The Helmholtz Machine (Dayan et al., 1995) presents a solution to this combinatorial explosion by maximizing an easily computed lower bound on the probability of the observations. The method can be viewed as a form of hierarchical self-supervised learning that may relate to feed-forward and feedback cortical pathways. Bottom-up "recognition" connections convert the input into representations in successive hidden layers, and top-down "generative" connections reconstruct the representation in one layer from the representation in the layer above. The network uses the inverse ("recognition") model to estimate the true posterior distribution of the input data.

Hinton (Hinton et al., 1995) proposed the "wake-sleep" algorithm for modifying the feedforward (recognition), and feedback (generative) weights of the Helmholtz machine. The "wake-sleep" algorithm employs the objective of "minimum description length" (Hinton and Zemel, 1994). The aim of learning is to minimize the total number of bits that would be required to communicate the input vectors by first sending the hidden unit representation, and then sending the difference between the input vector and the reconstruction from the hidden unit representation. Minimizing the description length forces the network to learn economical representations that capture the underlying regularities in the data.

A cost function C is defined as the total number of bits required to describe all of the hidden states in all of the hidden layers, α, plus the cost of describing the remaining information in the input vector d given the hidden states.

$$C(\alpha, d) = C(\alpha)C(d|\alpha) \tag{2.4}$$

The algorithm minimizes expected cost over all of the hidden states

$$E\left(C(\alpha, d)\right) = \sum_{\alpha} Q(\alpha|d)C(\alpha, d) \tag{2.5}$$

The conditional probability distribution over the hidden unit representations $Q(\alpha|d)$, needs to be estimated in order to compute the expected cost. The "wake-sleep" algorithm estimates $Q(\alpha|d)$ by driving the hidden unit activities via recognition connections from the input. These recognition connections are trained, in turn, by activating the hidden units and estimating the probability distributions of the input by generating "hallucinations" via the generative connections. Because the units are stochastic, repeating this process produces

may different hallucinations. The hallucinations provide an unbiased sample of the network's model of the world.

During the "wake" phase, neurons are driven by recognition connections, and the recognition model is used to define the objective function for learning the parameters of the generative model. The generative connections are adapted to increase the probability that they would reconstruct the correct activity vector in the layer below. During the "sleep" phase, neurons are driven by generative connections, and the generative model is used to define the objective function for learning the parameters of the recognition model. The recognition connections are adapted to increase the probability that they would produce the correct activity vector in the layer above.

The description length can be viewed as an upper bound on the negative log probability of the data given the network's generative model, so this approach is closely related to maximum likelihood methods of fitting models to data (Hinton et al., 1995). It can be shown that Bayesian inference models are equivalent to a minimum description length principle (Mumford, 1996). The generative models described in this section therefore fall under rubric of efficient coding. Another approach to the objective of efficient coding is explicit reduction of redundancy between units in the input signal. Redundancy can by minimized with the additional constraint on the number of coding units, as in minimum description length, or redundancy can be reduced without compressing the representation in a higher dimensional, sparse code.

1.2. Redundancy reduction as an organizational principle

Redundancy reduction has been proposed as a general organizational principle for unsupervised learning. Horace Barlow (Barlow, 1989) has argued that statistical redundancy contains information about the patterns and regularities of sensory stimuli. Completely non-redundant stimuli are indistinguishable from random noise, and Barlow claims that the percept of structure is driven by the dependencies. The set of points on the left of Figure 2.1 was selected randomly from a Gaussian distribution, whereas half of the points on the right were generated by rotating an initial set of points about the centroid of the distribution. This simple dependence between pairs of dots produced a structured appearance.

According to Barlow's theory, what is important for a system to detect is new statistical regularities in the sensory input that differ from the environment to which the system has been adapted. Barlow termed these new dependencies "suspicious coincidences." Bars and edges, for example, are locations in the visual input at which there is phase alignment across multiple spatial scales, and therefore constitute a "suspicious coincidence" (Barlow, 1994).

Learning mechanisms that encode the redundancy that is expected in the input and remove it from the output enable the system to more reliably detect

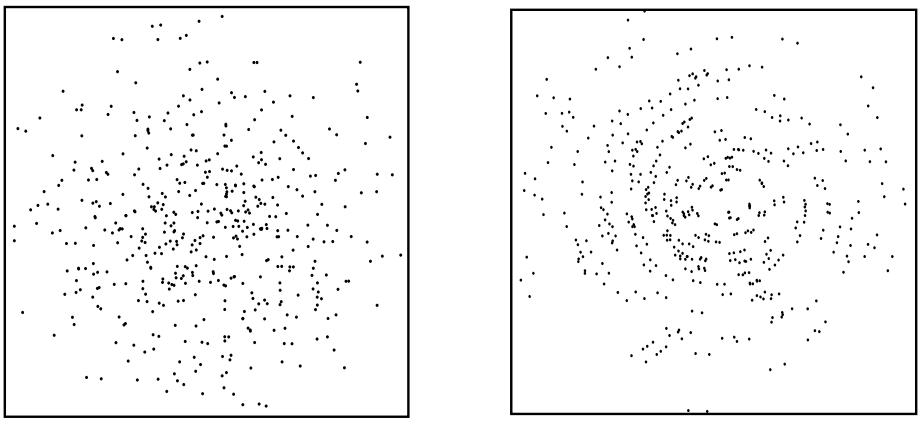

Figure 2.1. The percept of structure is driven by the dependencies. LEFT: A set of points selected from a Gaussian distribution. RIGHT: Half of the points were selected from a Gaussian distribution, and the other half were generated by rotating the points 5° about the centroid of the distribution. Figure inspired by Barlow (1989).

these new regularities. Learning such a transformation is equivalent to modeling the prior knowledge of the statistical dependencies in the input (Barlow, 1989). Independent codes are advantageous for encoding complex objects that are characterized by high order combinations of features because the prior probability of any particular high order combination is low. Incoming sensory stimuli are automatically compared against the null hypothesis of statistical independence, and suspicious coincidences signaling a new causal factor can be more reliably detected.

Barlow pointed to redundancy reduction at several levels of the visual system. Refer to Figure 2.2. A first-order redundancy is mean luminance. Adaptation mechanisms take advantage of this nonrandom feature by using it as an expected value, and expressing values relative to it (Barlow, 1989). The variance, a second-order statistic, is the luminance contrast. Contrast appears to be encoded relative to the local mean contrast, as evidenced by the "simultaneous contrast" illusion, and by contrast gain control mechanisms observed in V1 (Heeger, 1992).

1.3. Information theory

Barlow proposed an organizational principle for unsupervised learning based on information theory. The information provided by a given response x is defined as the number of bits required to communicate an event that has probability $P(x)$ under a distribution that is agreed upon by the sender and receiver (Shannon and Weaver, 1949):

$$I(x) = -log_2 P(x) \tag{2.6}$$

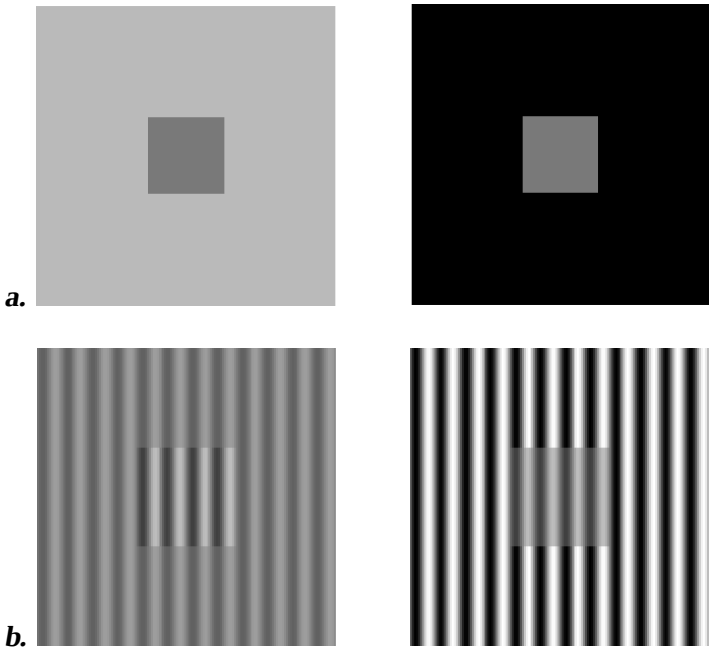

Figure 2.2. Redundancy reduction in the visual system. a. Luminance adaptation. The center squares are the same shade of gray, but the square on the left appears darker than the square on the right. b. Contrast adaptation. The center squares have the same contrast, but the square on the left appears to have higher contrast than the square on the right. This is called the simultaneous contrast effect.

Information is inversely proportional to the probability, and can be thought of as "surprise." The *entropy* of a response distribution, $H(x)$, is the expected value of the information:

$$H(x) = -\sum P(x) log_2 P(x) \tag{2.7}$$

Entropy is maximized by a uniform distribution, and is minimized by highly kurtotic (sharply peaked) distributions. The joint entropy between two variables x_1 and x_2 can be calculated as

$$H(x_1, x_2) = H(x_1) + H(x_2) - I(x_1, x_2) \tag{2.8}$$

where $I(x_1, x_2)$ is the mutual information between x_1 and x_2, which is calculated from 2.6 using the joint probability density $P(x_1, x_2)$.

Barlow argued for minimum entropy coding as a general representational strategy. Minimum entropy, highly kurtotic codes, have low mutual information between the elements. This is because the joint entropy of a multidimensional code is defined as the sum of the individual entropies minus the mutual information between the elements (2.8). Since the joint entropy of the code

stays constant, by minimizing the sum of the individual entropies, the mutual information term is also minimized. Another way to think of this is moving the redundancy from *between* the elements to redundancy *within* the distributions of the individual elements (Field, 1994). The distributions of individual elements with minimum entropy are redundant in the sense that they almost always take on the same value.

Atick and Redlich (Atick and Redlich, 1992) approach the objective of redundancy reduction from the perspective of efficient coding. They point out that natural stimuli are very redundant, and hence the sample of signals formed by an array of sensory receptors is inefficient. Atick (Atick, 1992) described evolutionary advantages of efficient coding such as coping with information bottlenecks due to limited bandwidth and limited dynamic range. Atick argued for the principle of efficiency of information representation as a design principle for sensory coding, and presented examples from the blowfly and the mammalian retina.

1.4. Redundancy reduction in the visual system

The large monopolar cells (LMC) in the blowfly compound eye eliminate inefficiency due to unequal use of neural response levels (Laughlin, 1981). The most efficient response gain is the one such that the probability distribution of the outputs is constant for all output states (maximum entropy). The solution is to match the gain of the transfer function to the cumulative probability density of the input. Laughlin (Laughlin, 1981) measured the cumulative probability density of contrast in the fly's environment, and found a close match between the gain of the LMC neurons and the cumulative probability density function.

Atick made a similar argument for the modulation transfer function (MTF) of the mammalian retina. The cumulative density of the amplitude spectrum of natural scenes is approximately $1/f$ where f is frequency[1] (Field, 1987). The MTF makes an efficient code by equalizing the response distribution of the output over spatial frequency. Atick demonstrated that multiplying the experimentally observed retinal MTF's by $1/f$ produces an approximately flat output for frequencies less than 3 cycles per degree. Atick refers to such transfer functions as whitening filters, since they equalize the response distribution of the output over all frequencies.

Macleod and von der Twer (Macleod and von der Twer, 1996) generalized Laughlin's analysis of optimal gain control to the presence of noise. In the noiseless case, the gain that maximizes the information transfer is the one that matches the cumulative probability density of the input, but in the presence of noise, the optimal transfer function has a shallower slope in order to increase

[1] Spatial frequency is determined by a Fourier transform on the wave form defined by brightness as a function of spatial position. In 2D images, a 1-D analysis is repeated at multiple orientations.

the signal-to-noise ratio. Macleod and von der Twer defined an optimal transfer function for color coding, which they termed the "pleistochrome," that maximizes the quantity of distinguishable colors in the presence of output noise. The analysis addressed the case of a single input x and output y, and used a criterion of minimum mean squared reconstruction error of the input, given the output plus output noise with variance σ. The minimum squared error criterion performs principal component analysis which, as will be discussed in the next section, maximizes the entropy of the output for the single unit case. In the presence of noise, the optimal transfer function was a gain proportional to $\sigma\left(P^{\frac{1}{3}}(x)\right)$, which was less than the cumulative probablility density, and modulated by the amount of noise, σ. Macleod and von der Twer found that the pleistochrome based on the distribution of cone responses along the $S-(L+M)$ axis[2] accounted well for the spectral sensitivity of the blue-yellow opponent channel observed at higher levels in the primate visual system.

These analyses have presented means for maximizing efficiency of coding for a single input and output. Principal component analysis is a means of reducing redundancies between multiple outputs. Atick and Redlich (Atick and Redlich, 1992) have argued for compact decorrelating mechanisms such as principal component analysis as a general coding strategy for the visual system. PCA decorrelates the input through an axis rotation. PCA provides a set of axes for encoding the input in fewer dimensions with minimum loss of information, in the squared error sense. Principal component analysis is an example of a coding strategy that in Barlow's formulation, encodes the correlations that are expected in the input and removes them from the output.

1.5. Principal component analysis

Principal component analysis (PCA) finds an orthonormal set of axes pointing in the directions of maximum covariance in the data. Let X be a dataset in which each column is an observation and each row is a measure with zero mean. The principal component axes are the eigenvectors of the covariance matrix of the measures, $\frac{1}{N}XX^T$, where N is the number of observations. The corresponding eigenvalues indicate the proportion of variability in the data for which each eigenvector accounts. The first principal component points in the direction of maximum variability, the second eigenvector points in the direction of maximum variability orthogonal to the first, and so forth. The data are recoded in terms of these axes by vector projection of each data point onto each of the new axes. Let P be the matrix containing the principal component eigenvectors in its columns. The PCA representation for each observation is

[2]Blue-yellow axis. S, M, and L stand for short, medium, and long wavelength selective cones. These correspond roughly to blue, green, and red. L+M corresponds to yellow.

obtained in the rows of A by

$$A = X^T P \tag{2.9}$$

The eigenvectors in P can be considered a set of weights on the data, X, where the outputs are the coefficients in the matrix, A. Because the principal component eigenvectors are orthonormal, they are also basis vectors for the dataset X. This is shown as follows: Since P is symmetric and the columns of P are orthonormal, $PP^T = I$, where I is the identity matrix, and right multiplication of 2.9 by P^T gives $AP^T = X$. The original data can therefore be reconstructed from the coefficients A using the eigenvectors in P now as basis vectors. A lower dimensional representation can be obtained by selecting a subset of the principal components with the highest eigenvalues, and it can be shown that for a given number of dimensions, the principal component representation minimizes mean squared reconstruction error.

Because the eigenvectors point in orthogonal directions in covariance space, the principal component representation is uncorrelated. The coefficients for one of the axes cannot be *linearly* predicted from the coefficients of the other axes. Another way to think about the principal component representation is in terms of the generative models described in Section 1.1. PCA models the data as a multivariate Gaussian where the covariance matrix is restricted to be diagonal. It can be shown that a generative model that maximizes the likelihood of the data given a Gaussian with a diagonal covariance matrix is equivalent to minimizing mean squared error of the generated data. PCA can also be accomplished through Hebbian learning, as described in the next section.

1.6. Hebbian learning

Hebbian learning is an unsupervised learning rule that was proposed as a model for activity dependent modification of synaptic strengths between neurons (Hebb, 1949). The learning rule adjusts synaptic strengths in proportion to the activity of the pre and post-synaptic neurons. Because simultaneously active inputs cooperate to produce activity in an output unit, Hebbian learning finds the correlational structure in the input. See (Becker and Plumbley, 1996) for a review of Hebbian learning.

For a single output unit, it can be shown that Hebbian learning maximizes activity variance of the output, subject to saturation bounds on each weight, and limits on the total connection strength to the output neuron (Linsker, 1988). Since the first principal component corresponds to the weight vector that maximizes the variance of the output, then Hebbian learning, subject to the constraint

that the weight vector has unit length, is equivalent to the finding first principal component of the input (Oja, 1982).

For a single output unit, y, where the activity of y is the weighted sum of the input, $y = \sum_i w_i x_i$, the simple Hebbian learning algorithm

$$\Delta w_i = \alpha x_i y \tag{2.10}$$

with learning rate α will move the vector $w = [w_1, ..., w_n]$ towards the first principal component of the input x. In the simple learning algorithm, the length of w is unbounded. Oja modified this algorithm so that the length of w was normalized after each step. With a sufficiently small α, Hebbian learning with length normalization is approximated by

$$\Delta w = \alpha y(x - wy). \tag{2.11}$$

This learning rule converges to the unit length principal component. The $-wy^2$ term tends to decease the length of w if it gets too large, while allowing it to increase if it gets too small.

In the case of N output units, in which the N outputs are competing for activity, Hebbian learning can span the space of the first N principal components of the input. With the appropriate form of competition, the Hebb rule explicitly represents the N principal components in the activities of the output layer (Oja, 1989; Sanger, 1989). A learning rule for the weight w_j to output unit y_j that explicitly finds the first N principal components of the data is

$$\Delta w_j = \alpha y_j (x - w_j y_j + 2 \sum_{k=1}^{j-1}) \tag{2.12}$$

The algorithm forces successive outputs to learn successive principal components of the data by subtracting estimates of the previous components from the input before the connections to a given output unit are updated.

Linsker (Linsker, 1988) also demonstrated that for the case of a single output unit, Hebbian learning maximizes the information transfer between the input and the output. The Shannon information transfer rate

$$R = I(x, y) = H(y) - H(y|x) \tag{2.13}$$

gives the amount of information that knowing the output y conveys about the input x, and is equivalent to the mutual information between them, $I(x, y)$. For a single output unit y with a Gaussian distribution, 2.13 is maximized by maximizing the variance of the output (Linsker, 1988). Maximizing output variance within the constraint of a Gaussian distribution produces a response distribution that is as flat as possible (i.e. high entropy). Maximizing output

entropy with respect to a weight w maximizes 2.13, because the second term, $H(y|x)$, is noise and does not depend on w.

Linsker argued for maximum information preservation as an organizational principle for a layered perceptual system. There is no need for any higher layer to attempt to reconstruct the raw data from the summary received from the layer below. The goal is to preserve as much information as possible in order to enable the higher layers to use environmental information to discriminate the relative value of different actions. In a series of simulations described later in this chapter, in Section 3, Linsker (Linsker, 1986) demonstrated how structured receptive fields[3] with feature-analyzing properties related to the receptive fields observed in the retina, LGN, and visual cortex could emerge from the principle of maximum information preservation. This demonstration was implemented using a local learning rule[4] subject to constraints. Information maximization has recently been generalized to the multi-unit case (Bell and Sejnowski, 1995). Information maximization in multiple units will be discussed below in Section 2. This monograph examines representations for face images based on information maximization.

1.7. Learning rules for explicit discovery of statistical dependencies

A perceptual system can be organized around internally derived teaching signals generated from the assumption that different parts of the perceptual input have common causes in the external world. One assumption is that the visual input is derived from physical sources that are approximately constant over space. For example, depth tends to vary slowly over most of the visual input except at object boundaries. Learning algorithms that explicitly encode statistical dependencies in the input attempt to discover those constancies. The actual output of such invariance detectors represents the extent to which the current input violates the network's model of the regularities in the world (Becker and Plumbley, 1996). The Hebbian learning mechanism described in the previous section is one means for encoding the second order dependencies (correlations) in the input.

The GMAX algorithm (Pearlmutter and Hinton, 1986) is a learning rule for multiple inputs to a single output unit that is based on the goal of redundancy reduction. The algorithm compares the response distribution, P of the output unit to the response distribution, Q, that would be expected if the input was

[3] A receptive field of a neuron is the input that influences its activity rate. Many neurons in the retina and lateral geniculate nucleus of the thalamus (LGN) have receptive fields with excitatory centers and inhibitory surrounds. These respond best to a spot of light surrounded by a dark annulus at a particular location in the visual field. Many neurons in the primary visual cortex respond best to oriented bars or edges.

[4] Local learning rules may be more biologically plausible than rules that evaluate information from all units, given the limited extent of synaptic connections

entirely independent. The learning algorithm causes the unit to discover the statistical dependencies in the input by maximizing the difference between P and Q. P is determined by the responses to the full set of data under the current weight configuration, and Q can be calculated explicitly by sampling all of the 2^n possible states of the n input units. The GMAX learning rule is limited to the case of a single output unit, and probabilistic binary units.

Becker (Becker, 1992) generalized GMAX to continuous inputs with Gaussian distributions. This resulted in a learning rule that minimized the ratio of the output variance to the variance that would be expected if the input lines were independent. This learning rule discovers statistical dependencies in the input, and is literally an invariance detector. If we assume that properties of the visual input are derived from constant physical sources, then a learning rule that minimizes the variance of the output will tell us something about that physical source. Becker further generalized this algorithm to the case of multiple output units. These output units formed a mixture model of different invariant properties of the input patterns.

Becker and Hinton (Becker and Hinton, 1992; Becker and Hinton, 1993) applied the multi-unit version of this learning rule to show how internally derived teaching signals for a perceptual system can be generated from the assumption that different parts of the perceptual input have common causes in the external world. In their learning scheme, small modules that look at separate but related parts of the perceptual input discover these common causes by striving to produce outputs that agree with each other. The modules may look at different modalities such as vision and touch, or the same modality at different times, such as the consecutive two-dimensional views of a rotating three-dimensional object, or spatially adjacent parts of the same image. The learning rule, which they termed IMAX, maximizes the mutual information between pairs of output units, y_a and y_b. Under the assumption that the two output units are caused by a common underlying signal corrupted by independent Gaussian noise, then the mutual information between the underlying signal and the mean of y_a and y_b is given by

$$I = 0.5 \;\; log \frac{V(y_a + y_b)}{V(y_a - y_b)} \tag{2.14}$$

where V is the variance function over the training cases. The algorithm can be understood as follows: A simple way to make the outputs of the two modules agree is to use the squared difference between the module outputs as a cost function (the denominator of 2.14). A minimum squared difference cost function alone, however will cause both modules to produce the same constant output that is unaffected by the input, and therefore convey no information about the input. The numerator modified the cost function to minimize the squared difference relative to how much both modules varied as the input varied. This

forced the modules to respond to something that was common in their two inputs.

Becker and Hinton showed that maximizing the mutual information between spatially adjacent parts of an image can discover depth in random dot stereograms of curved surfaces. The simulation consisted of a pair of 2-layer networks, each with a single output unit, that took spatially distinct regions of the visual space as input. The input consisted of random dot stereograms with smoothly varying stereo disparity. Following training, the module outputs were proportional to depth, despite no prior knowledge of the third dimension. The model was extended to develop population codes for stereo disparity (Becker and Hinton, 1992), and to model the locations of discontinuities in depth (Becker, 1993).

Schraudolph and Sejnowski (Schraudolph and Sejnowski, 1992) proposed an algorithm for learning invariances that was closely related to Becker and Hinton's constrained variance minimization. They combined a variance-minimizing anti-Hebbian term, in which connection strengths are *reduced* in proportion to the pre-and post synaptic unit activities, with a term that prevented the weights from converging to zero. They showed that a set of competing units could discover population codes for stereo disparity in random dot stereograms.

Zemel and Hinton (Zemel and Hinton, 1991) applied the IMAX algorithm to the problem of learning to represent the viewing parameters of simple objects, such as the object's scale, location, and size. The algorithm attempts to learn multiple features of a local image patch that are uncorrelated with each other, while being good predictors of the feature vectors extracted from spatially adjacent input locations. The algorithm is potentially more powerful than linear decorrelating methods such as principal component analysis because it combines the objective of decorrelating the feature vector with the objective of finding common causes in the spatial domain. Extension of the algorithm to more complex inputs than synthetic 2-D objects is limited, however, due to the difficulty of computing the determinants of ill-conditioned matrices (Becker and Plumbley, 1996).

2. INDEPENDENT COMPONENT ANALYSIS

2.1. Decorrelation versus independence

Principal component analysis *decorrelates* the input data, but does not address the high-order dependencies. Decorrelation simply means that variables cannot be predicted from each other using a *linear* predictor. There can still be nonlinear dependencies between them. Consider two variables, x and y that are related to each other by a sine wave function, $y = sin(x)$. The correlation coefficient for the variables x and y would be zero, but the two variables are highly dependent nonetheless. Edges, defined by phase alignment at multiple

spatial scales, are an example of a high-order dependency in an image, as are elements of shape end curvature.

Second-order statistics capture the amplitude spectrum of images but not the phase (Field, 1994). Amplitude is a second-order statistic. The amplitude spectrum of a signal is essentially a series of correlations with a set of sine-waves. Also, the Fourier transform of the autocorrelation function of a signal is equal to its power spectrum (square of the amplitude spectrum). Hence the amplitude spectrum and the autocorrelation function contain the same information. The remaining information that is not captured by the autocorrelation function, the high order statistics, corresponds to the phase spectrum.[5]

Coding mechanisms that are sensitive to phase are important for organizing a perceptual system. Spatial phase contains the structural information in images that drives human recognition much more strongly than the amplitude spectrum (Oppenheim and Lim, 1981; Piotrowski and Campbell, 1982). For example, A face image synthesized from the amplitude spectrum of face A and the phase spectrum of face B will be perceived as an image of face B.

Independent component analysis (ICA) (Comon, 1994) is a generalization of principal component analysis that separates the high-order dependencies in the input, in addition to the second-order dependencies. As noted above, principal component analysis is a way of encoding second order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. Consider a set of data points derived from two underlying distributions as shown in Figure 2.3. Principal component analysis models the data as a multivariate Gaussian and would place an orthogonal set of axes such that the two distributions would be completely overlapping. Independent component analysis does not constrain the axes to be orthogonal, and attempts to place them in the directions of maximum statistical dependencies in the data. Each weight vector in ICA attempts to encode a portion of the dependencies in the input, so that the dependencies are removed from between the elements of the output. The projection of the two distributions onto the ICA axes would have less overlap, and the output distributions of the two weight vectors would be kurtotic (Field, 1994).[6] Algorithms for finding the independent components of arbitrary data sets are described in Section 2.2

2.2. Information maximization learning rule

Bell and Sejnowski (Bell and Sejnowski, 1995) recently developed an algorithm for separating the statistically independent components of a dataset through unsupervised learning. The algorithm is based on the principle of

[5] Given a translation invariant input, it is not possible to compute any statistics of the phase from the amplitude spectrum (Dan Ruderman, personal communication.)

[6] Thanks to Michael Gray for this observation.

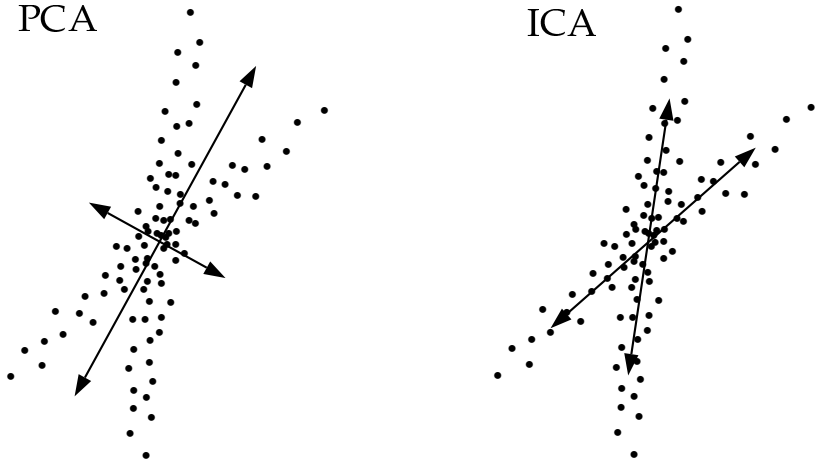

Figure 2.3. Example 2-D data distribution and the corresponding principal component and independent component axes. The data points could be, for example, grayvalues at pixel 1 and pixel 2. Figure inspired by Lewicki & Sejnowski (2000).

maximum information transfer between sigmoidal neurons. This algorithm generalizes Linsker's information maximization principle (Linsker, 1988) to the multi-unit case and maximizes the joint entropy of the output units. Another way of describing the difference between PCA and ICA is therefore that PCA maximizes the joint *variance* of the outputs, whereas ICA maximizes the joint *entropy* of the outputs.

Bell and Sejnowski's algorithm is illustrated as follows: Consider the case of a single input, x, and output, y, passed through a nonlinear squashing function:

$$u = wx + w_0 \qquad y = g(u) = \frac{1}{1 + e^{-u}}. \tag{2.15}$$

As illustrated in Figure 2.4, the optimal weight w on x for maximizing information transfer is the one that best matches the probability density of x to the slope of the nonlinearity. The optimal w produces the flattest possible output density, which in other words, maximizes the entropy of the output.

The optimal weight is found by gradient ascent on the entropy of the output, y with respect to w:

$$\frac{\partial}{\partial w} H(y) = \frac{\partial}{\partial w} - \sum P(y) log_2 P(y). \tag{2.16}$$

Maximizing the entropy of the output is equivalent to maximizing the mutual information between the input and the output (i.e. maximizing information

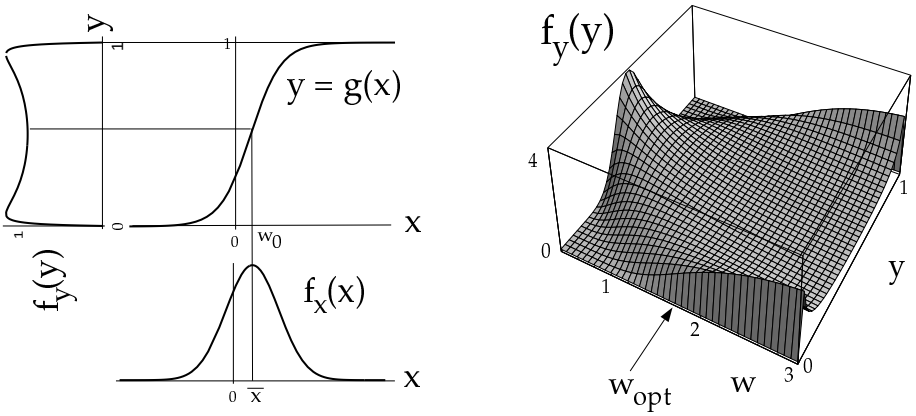

Figure 2.4. Optimal information flow in sigmoidal neurons. The input x is passed through a nonlinear function, $g(x)$. The information in the output density $f_y(y)$ depends on matching the mean and variance of $f_x(x)$ to the slope and threshold of $g(x)$. Right: $f_y(y)$ is plotted for different values of the weight, w. The optimal weight, w_{opt} transmits the most information. Figure from Bell & Sejnowski (1995), reprinted with permission from MIT Press, copyright 1995, MIT Press.

transfer). This is because $I(x,y) = H(x) + H(y) - H(y|x)$, where only $H(y)$ depends on the weight w since $H(y|x)$ is noise.

When there are multiple inputs and outputs, $X = (x_1, x_2, ...)$, $Y = (y_1, y_2, ...)$ maximizing the joint entropy of the output encourages the individual outputs to move towards statistical independence. To see this, we refer back to Equation 2.8: $H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2)$. Maximizing the joint entropy of the output $H(y_1, y_2, ...)$ encourages the mutual information between the individual outputs $I(y_1, y_2, ...)$ to be small. The mutual information is guaranteed to reach a minimum when the nonlinear transfer function g matches the cumulative distribution of the independent signals responsible for the data in x, up to scaling and translation (Nadal and Parga, 1994; Bell and Sejnowski, 1997). Many natural signals, such as sound sources, have been shown to have a super-Gaussian distribution, meaning that the kurtosis of the probability distribution exceeds that of a Gaussian (Bell and Sejnowski, 1995). For mixtures of super-Gaussian signals, the logistic transfer function has been found to be sufficient to separate the signals (Bell and Sejnowski, 1995).

Since $y = g(x)$ and g is monotonic, the probability $P(y)$ in Equation 2.16 can be written in terms of $P(x)$ in the single unit case as (Papoulis, 1991)

$$P(y) = \frac{P(x)}{\frac{\partial y}{\partial x}} \text{ and in the multiunit case as } P(Y) = \frac{P(X)}{|J|}$$

where $|J|$ is the determinant of the Jacobian, J. J is the matrix of partial derivatives $\frac{\partial y_i}{\partial x_i}$. Hence

$$H(Y) = -E\left(log_2 \frac{P(X)}{|J|}\right) = H(X) + E\left(log_2 |J|\right). \quad (2.17)$$

Since $H(X)$ does not depend on W, the problem reduces to maximizing $|J|$ with respect to W. Computing the gradient of $|J|$ with respect to W results in the following learning rule:[7]

$$\Delta W = \alpha \left((W^T)^{-1} + y' x^T\right) \quad (2.18)$$

$$\text{where } y' = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i} = \frac{\partial}{\partial u_i} ln \frac{\partial y_i}{\partial u_i}.$$

Bell & Sejnowski improved the algorithm in 1997 by using the natural gradient (Amari et al., 1996). They multiplied the gradient equation by the symmetric matrix $W^T W$ which removed the inverse and scaled the gradient differently along different dimensions. The natural gradient addresses the problem that the metric space of W is not necessarily Euclidean. Each dimension has its own scale and the natural gradient normalizes the metric function for that space. This resulted in the following learning rule:

$$\Delta W = \alpha(I + y' x^T W^T) W \quad (2.19)$$

Although it appears at first contradictory, information maximization in a multidimensional code is consistent with Barlow's notion of minimum entropy coding. Refer again to Equation 2.8. As noted above, maximizing the *joint* entropy of the output encourages the mutual information between the outputs to be small, but under some conditions other solutions are possible for which the mutual information is nonzero. Given that the joint entropy stays constant (at its maximum), the solution that minimizes the mutual information will also minimize the *marginal* (individual) entropies of the output units.

An application of independent component analysis is signal separation. Mixtures of independent signals can be separated by a weight matrix that minimizes the mutual information between the outputs of the transformation. Bell & Sejnowski's information maximization algorithm successfully solved the "cocktail party" problem, in which a set random mixtures of auditory signals were separated without prior knowledge of the original signals or the mixing process (Bell and Sejnowski, 1995). The algorithm has also been applied to separating the sources of EEG signals (Makeig et al., 1996), and fMRI images (McKeown et al., 1998).

Independent component analysis can be considered as a generative model of the data assuming independent sources. Each data point x is assumed to be a linear mixture of independent sources, $x = As$, where A is a mixing matrix, and

[7] The step from 2.17 to 2.18 is presented in the Appendix of (Bell and Sejnowski, 1995).

s contains the sources. Indeed, a maximum likelihood approach for finding A and s can be shown to be mathematically equivalent to the information maximization approach of Bell and Sejnowski (MacKay, 1996; Pearlmutter and Parra, 1996). In the maximum likelihood approach, a likelihood function of the data is generated under the model $x = As$, where the probabilities of the sources s are assumed to be factorial. The elements of the basis matrix A and the sources s are then obtained by gradient ascent on the log likelihood function. Another approach to independent component analysis involves cost functions using marginal cumulants (Comon, 1994; Cardoso and Laheld, 1996). The adaptive methods in the information maximization approach are more plausible from a neural processing perspective than the cumulant-based cost functions (Lee, 1998).

A large variety of algorithms have been developed to address issues including extending the information maximization approach to handle sub-Gaussian sources (Lee et al., 1999), estimating the shape of the distribution of input sources with maximum likelihood techniques (Pearlmutter and Parra, 1996), nonlinear independent component analysis (Yang et al., 1998), and biologically inspired algorithms that perform ICA using local computations (Lin et al., 1997). I refer you to (Lee, 1998) for a thorough review of algorithms for independent component analysis.

2.3. Relation of sparse coding to independence

Atick argued for compact, decorrelated codes such as PCA because of efficiency of coding. Field (Field, 1994) argued for sparse, distributed codes in favor of such compact codes. Sparse representations are characterized by highly kurtotic response distributions, in which a large concentration of values are near zero, with rare occurrences of large positive or negative values in the tails. Recall that highly kurtotic response distributions have low entropy. Maximizing sparseness of a response distribution is related to minimizing its entropy, and sparse codes therefore incur the same advantages as minimum entropy codes, such as separation of high-order redundancies in addition to the second-order redundancy. In such a code, the redundancy *between* the elements of the input is transformed into redundancy *within* the response patterns of the individual outputs, where the individual outputs almost always give the same response except on rare occasions.

Given this relationship between sparse codes and minimum entropy, the advantages of sparse codes as outlined in (Field, 1994) are also arguments in favor of Barlow's minimum entropy codes (Barlow, 1989). Codes that minimize the number of active neurons can be useful in the detection of suspicious coincidences. Because a nonzero response of each unit is relatively rare, high order relations become increasingly rare, and therefore more informative when they are present in the stimulus. Field contrasts this with a compact code

such as principal components, in which a few cells have a relatively high probability of response, and therefore high order combinations among this group are relatively common. In a sparse distributed code, different objects are represented by *which* units are active, rather than by their *rate* of activity. These representations have an added advantage in signal-to-noise, since one need only determine which units are active without regard to the precise level of activity. An additional advantage of sparse coding for face representations is storage in associative memory systems. Networks with sparse inputs can store more memories and provide more effective retrieval with partial information (Palm, 1980; Baum et al., 1988).

Field presented evidence that oriented Gabor filters produce sparse codes when presented with natural scenes, whereas the response distribution is Gaussian when presented with synthetic images generated from $1/f$ noise. Because the two image classes had the same amplitude spectra and differed only in phase, Field concluded that sparse coding by Gabor filters depends primarily on the phase spectra of the data. Olshausen and Field (Olshausen and Field, 1996b; Olshausen and Field, 1996a) showed that a generative model with a sparseness objective can account for receptive fields observed in the primary visual cortex. They trained a network to reconstruct natural images from a linear combination of unknown basis images with minimum mean-squared error. The minimum squared error criterion alone would have converged on a linear combination of the principal components of the images. When a sparseness criterion was added to the objective function, the learned basis images were local, oriented, and spatially opponent, similar to the response properties of V1 simple cells.[8] Maximizing sparseness under the constraint of information preservation is equivalent to minimum entropy coding.

Bell & Sejnowski also examined an image synthesis model of natural scenes using independent component analysis (Bell and Sejnowski, 1997). As expected given the relationship between sparse coding and independence, Bell & Sejnowski obtained a similar result to Olshausen and Field, namely the emergence of local, spatially opponent receptive fields. Moreover, the response distributions of the individual output units were indeed sparse. Decorrelation mechanisms such as principal components resulted in spatially opponent receptive fields, some of which were oriented, but were not spatially local. In addition, the response distributions of the individual PCA output units were Gaussian. In a related study, Wachtler, Lee, and Sejnowski (Wachtler et al., 2001) performed ICA on chromatic images of natural scenes. Redundancy reduction was much higher in the chromatic case than in the grayscale case. The

[8] "Simple cells" in the primary visual cortex respond to an oriented bar at a precise location in the visual field. There is a surrounding inhibitory region, such that the receptive field is similar to a sine wave grating modulated by a Gaussian.

resulting filters segmented into color opponent and broadband filters, paralleling the color opponent and broadband channels in the primate visual system. These filters had very sparse distributions, suggesting that color opponency in the human visual system achieves a highly efficient representation of colors.

3. UNSUPERVISED LEARNING IN VISUAL DEVELOPMENT

3.1. Learning input dependencies: Biological evidence

There is a large body of evidence that self-organization plays a considerable role in the development of the visual system, and that this self-organization is mediated by learning mechanisms that are sensitive to dependencies in the input. The gross organization of the visual system appears to be governed by molecular specificity mechanisms during embryogenesis (Harris and Holt, 1990). Such processes as the generation of the appropriate numbers of target neurons, migration to the appropriate position, the outgrowth of axons, their navigation along appropriate pathways, recognition of the target structure, and the formation of at least coarsely defined topographic maps[9] may be mediated by molecular specificity. During postnatal development, the architecture of the visual system continues to become defined, organizing into ocular dominance and orientation columns.[10] The statistical properties of early visual experience and endogenous activity appear to be responsible for shaping this architecture. See (Stryker, 1991a) for a review.

Learning mechanisms that are sensitive to dependencies in the visual input transform these statistical properties into cortical receptive field architecture. The NMDA receptor could be the "correlation detector" for Hebbian learning between neurons. It opens calcium channels in the post synaptic cell in a manner that depends on activity in both the pre- and the post-synaptic cell. Specifically, it depends on glutamate from the presynaptic cell and the voltage of the post synaptic cell. Although it is not known exactly how activation of the NMDA receptor would lead to alterations in synaptic strength, several theories have been put forward involving the release of trophic substances, retrograde messenger systems leading back to the presynaptic neuron, and synaptic morphology changes (Rison and Stanton, 1995).

Visual development appears to be closely associated with NMDA gating (Constantine-Paton et al., 1990). There is longer NMDA gating during visual development, which provides a longer temporal window for associations. Levels of NMDA are high early in development, and then drop (Carmignoto and

[9] Neighboring neurons tend to respond to neighboring regions of the visual field.

[10] Adjacent neurons in the primary visual cortex prefer gradually varying orientations. Perpendicular to this are iso-orientation stripes. Eye preference is also organized into stripes.

Vicini, 1992). These changes in NMDA activity appear to be dependent on experience rather than age. Dark rearing will delay the drop in NMDA levels, and the decrease in length of NMDA gating is also dependent on activity (Fox et al., 1992).

The organization of ocular dominance and orientation preference can be altered by manipulating visual experience. Monocular deprivation causes a greater proportion of neurons to prefer the active eye at the expense of the deprived eye (Hubel et al., 1977). Colin Blakemore (Blakemore, 1991) found that in kittens reared in an environment consisting entirely of vertical stripes, orientation preference in V1 was predominantly vertical. The segregation of ocular dominance columns is dependent on both pre- and post-synaptic activity. Ocular dominance columns do not form when all impulse activity in the optic nerve is blocked by injecting tetrodotoxin (Stryker and Harris, 1986). Blocking post-synaptic activity during monocular deprivation nulls the usual shift in ocular dominance (Singer, 1990; Gu and Singer, 1993). Stryker demonstrated that ocular dominance segregation depends on asynchronous activity in the two eyes (Stryker, 1991a). With normal activity blocked, Stryker stimulated both optic nerves with electrodes. When the two nerve were stimulated synchronously, ocular dominance columns did not form, but when they were stimulated asynchronously, columns did form. Consistent with the role of NMDA in the formation of ocular dominance columns, NMDA receptor antagonists prevented the formation of ocular dominance columns, whereas increased levels of NMDA sharpened ocular dominance columns (Debinski et al., 1990). Some of organization of ocular dominance and orientation preference does occur prenatally. Endogenous activity can account for the segregation of ocular dominance in the lateral geniculate nucleus (Antonini and Stryker, 1993), and endogenous activity tends to be correlated in neighboring retinal ganglion cells (Mastronarde, 1989).

Intrinsic horizontal axon collaterals in the striate cortex of adult cats specifically link columns having the same preferred orientation. Calloway and Katz (Callaway and Katz, 1991) demonstrated that the orientation specificity of these horizontal connections was dependent on correlated activity from viewing sharply oriented visual stimuli. Crude clustering of horizontal axon collaterals is normally observed in the striate cortex of kittens prior to eye opening. Binocular deprivation beyond this stage dramatically affected the refinement of these clusters. Visual experience appears to have been necessary for adding and eliminating collaterals in order to produce the sharply tuned specificity normally observed in the adult.

3.2. Models of receptive field development based on correlation sensitive learning mechanisms

Orientation columns are developed prenatally in macaque. Therefore any account of their development must not depend on visual experience. Linsker (Linsker, 1986) demonstrated that orientation columns can arise from random input activity in a layered system using Hebbian learning. The only requirements for this system were arborization functions that were more dense centrally, specification of initial ratios of excitatory and inhibitory connections, and adjustment of parameters controlling the total synaptic strength to a unit. Because of the dense central connections, the random activity in the first layer became locally correlated in the second layer. Manipulation of the parameter for total synaptic strength in the third layer brought on center-surround receptive fields. This occurred because of the competitive advantage of the dense central connections over the sparse peripheral connections. Activity in the central region became saturated first, and because of the bounds on activity, the peripheral region became inhibitory. The autocorrelation function for activity in layer 3 was Mexican hat shaped. Linsker added four more layers to the network. The first three of these layers also developed center-surround receptive fields. The effect of adding these layers was to sharpen the Mexican hat autocorrelation function with each layer. Linsker associated the four center-surround layers of his model to the bipolar, retinal ganglion, LGN, and layer 4c cells in the visual system. A criticism of this section of Linsker's model is that it predicts that the autocorrelation function in these layers should become progressively more sharply Mexican hat shaped, which does not appear to occur in the primate visual system.

In the next layers of the model, Linsker demonstrated the development of orientation selective cells and their organization into orientation columns. Cells receiving inputs with a Mexican hat shaped autocorrelation function attempted to organize their receptive fields into banded excitatory and inhibitory regions. By adjusting the parameter for total synaptic strength in layer seven, Linsker was able to generate oriented receptive fields. Linsker subsequently generated iso-orientation bands by adding lateral connections in the top layer. The lateral connections were also updated by a Hebbian learning rule. Activity in like-oriented cells is correlated when the cells are aligned along the axis of orientation preference, but are anticorrelated on an axis perpendicular to the preferred orientation. The lateral connections thus encourage the same orientation along the axis of preferred orientation, and an orthogonal orientation preferences along the axis orthogonal to the preferred orientation. This organization resembles the singularities in orientation preference reported by Obermayer and Blasdel (Obermayer and Blasdel, 1993). In Linsker's model,

a linear progression of orientation preference would require an isotropic autocorrelation function.

Miller, Keller, and Stryker (Miller et al., 1989) demonstrated that Hebbian learning mechanisms can account for the development of ocular dominance slabs and for experience-related alterations of this organization. In their model, synaptic strength was altered as a function of pre and post synaptic activity, where synaptic strength depended on within-eye and between-eye correlation functions. The model also contained constraints on the overall synaptic strength, an arborization function indicating the initial patterns of connectivity, and lateral connections between the cortical cells. All input connections were excitatory.

Miller et al. found that there were three conditions necessary for the development of ocular dominance columns. 1. The input activity must favor monocularity by having larger within-eye correlations than between-eye correlations. 2. There must be locally excitatory cortical connections. 3. If the intracortical connections are not Mexican hat shaped, in other words if they do not have an inhibitory zone, then there must be a constraint on the total synaptic strength of the afferent axons. The ocular dominance stripes arose because of the intracortical activation function. If this function is Mexican hat shaped, then each cell will want to be in an island of like ocularity surrounded by opposite ocularity. Optimizing this force along a surface of cells results in a banded pattern of ocular dominance. The intracortical activation function controls the periodicity of the stripes. The ocular dominance stripes will have a periodicity equal to the fundamental frequency of the intracortical activation function. This will be the case up to the limit of the arborization function. If the excitatory region of the intracortical activation function is larger than the arborization function, then the periodicity of the stripes will be imposed by the arborization function.

Miller et al. found that a very small within-eye correlation function was sufficient to create ocular dominance stripes, so long as it was larger than the between eye correlation. Anticorrelation within an eye decreases monocularity, whereas anticorrelation between eyes, such as occurs in conditions of strabismus and monocular deprivation, increases monocularity. They also observed an effect related to critical periods. Monocular cells would remain stabilized once formed, and binocular cells would also stabilize if the synapses were at saturating strength. Therefore, alterations could only be made while there were still binocular cells with unsaturated connections. Due to the dependence of ocular dominance on excitatory intracortical connections, their simulation

predicted that ocular dominance organization in the developing brain would be eliminated by increasing inhibition[11].

Berns, Dayan, and Sejnowski (Berns et al., 1993) presented a Hebbian learning model for the development of both monocular and binocular populations of cells. The model is driven by correlated activity in retinal ganglion cells within each eye before birth, and between eyes after birth. An initial phase of same-eye correlations, followed by a second phase that included correlations between the eyes produced a relationship between ocular dominance and disparity that has been observed in the visual cortex of the cat. The binocular cells tended to be selective for zero disparity, whereas the more monocular cells tended to have nonzero disparity.

Obermayer, Blasdel, and Schulten (Obermayer et al., 1992) modeled the simultaneous development of ocular dominance and orientation columns with a Kohonen self-organizing topographic map. This algorithm predicts the observed geometrical relations between ocular dominance and orientation preference on the surface of the primary visual cortex. These include the perpendicular iso-orientation slabs in the binocular regions, and singularities in orientation preference at the centers of highly monocular zones. According to their model, cortical geometry is a result of projecting five features onto a two dimensional surface. The five features are spatial position along the horizontaland vertical axes, orientation preference, orientation specificity, and ocular dominance. The Kohonen self organizing map operates in the following way. The weights of the network attempt to learn a mapping from a five dimensional input vector onto a 2-D grid. The weight associated with each point on the grid is the combination of the five features preferred by that unit. The unit with the most similar weight vector to a given input vector, as measured by the dot product, adjusts its weight vector toward the input vector. Neighboring units on the grid also learn by a smaller amount according to a neighborhood function. At the beginning of training, the "temperature" is set to a high level, meaning that the neighborhood function is broad and the learning rate is high. The temperature is gradually reduced during training. The overall effect of this procedure is to force units on the grid to vary their preferences smoothly and continuously, subject to the input probabilities. Like Hebbian learning, the self organizing map creates structure from the correlations in input patterns, but the self organizing map has the added feature that the weights are forced to be smooth and continuous over space.

Obermayer, Blasdel, and Schulten likened the development of cortical geometry to a Markov random process. There are several possible states of cortical geometry, and the statistical structure of the input vectors trigger the transitions between states. They showed that a columnar system will not develop if the

[11] e.g. through application of muscimol, A GABA agonist, where GABA is an inhibitory neurotransmitter

input patterns are highly similar with respect to orientation preference, specificity, and ocular dominance. Nor will it segregate into columns if the inputs are entirely uncorrelated. There is a range of input correlations for which columnar organization will appear. Their model predicts that ocular dominance and orientation columns will be geometrically unrelated in animals that are reared with an orientation bias in one eye.

4. LEARNING INVARIANCES FROM TEMPORAL DEPENDENCIES IN THE INPUT

The input to the visual system contains not only spatial redundancies, but temporal redundancies as well. There are several synaptic mechanisms that might depend on the correlation between synaptic input at one moment, and post-synaptic depolarization at a later moment. Coding principles that are sensitive to temporal as well as spatial redundancies in the input may play a role in learning constancies of the environment such as viewpoint invariances.

Internally driven teaching signals can be derived not only from the assumption that *spatially* distinct parts of the perceptual input have common causes in the external world, but also from the assumption that *temporally* distinct inputs can have common causes. Objects have temporal persistence. They do not simply appear and disappear. Different views of an object or face tend to appear in close temporal proximity as an animal manipulates the object or navigates around it, or as a face changes expression or pose. Capturing the temporal relationships in the input is a way to associate different views of an object, and thereby learn representations that are invariant to changes in viewpoint.

4.1. Computational models

Földíak (Földiák, 1991) demonstrated that Hebbian learning can capture temporal relationships in a feedforward system when the output unit activities undergo temporal smoothing. Hebbian learning strengthens the connections between simultaneously active units. With the lowpass temporal filter on the output unit activities, Hebbian learning strengthens the connections between active inputs and *recently* active outputs. As discussed in Section 1.5, competitive Hebbian learning can find the principal components of the input data. Incorporating a hysteresis in the activation function allows competitive Hebbian mechanisms to find the spatio-temporal principal components of the input.

Peter Földíak (Földiák, 1991) used temporal association to model the development of translation independent orientation detectors such as the complex cells[12] of V1. His model was a two-layer network in which the input layer con-

[12] Unlike "simple cells", a "complex cell" in primary visual cortex is excited by a bar of a particular orientation at *any location* within its receptive field.

sisted of sets of local position dependent orientation detectors. This layer was fully connected to four output units. Földíak modified the traditional Hebbian learning rule such that weight changes would be proportional to presynaptic activity and a trace (running average) of postsynaptic activity. The network was trained by sweeping one orientation at a time across the entire input field such as may occur during prenatal development (Mastronarde, 1989; Meister et al., 1991). One representation unit would become active due to the competition in that layer, and it would stay active as the input moved to a new location. Thus units signaling "horizontal" at multiple locations would strengthen their connections to the same output unit that would come to represent "horizontal" at any location.

This mechanism can learn viewpoint-tolerant representations when different views of an object are presented in temporal continuity (Földiák, 1991; Weinshall and Edelman, 1991; Rhodes, 1992; O'Reilly and Johnson, 1994; Wallis and Rolls, 1997). Földíak achieved translation invariance in a single layer by having orientation-tuned filters in the first layer that provided linearly separable patterns to the next layer. More generally, approximate viewpoint invariance may be achieved by the superposition of several Földíak-like networks (Rolls, 1995).

O'Reilly and Johnson (O'Reilly and Johnson, 1994) modeled translation invariant object recognition based on reciprocal connections between layers and lateral inhibition within layers. Their architecture was based on the anatomy of the chick IMHV, a region thought to be involved in imprinting. In their model, the reciprocal connections caused a hysteresis in the activity of all of the units, which allowed Hebbian learning to associate temporally contiguous inputs. The model demonstrated that a possible function of reciprocal connections in visual processing areas is to learn translation invariant object recognition. The model also suggested an interpretation of critical periods. Chicks are only able to imprint new objects early in development. As an object was continuously presented to the network, more and more units were recruited to represent that object. Only unrecruited units and units without saturated connections could respond to the new objects.

Becker (Becker, 1993) showed that the IMAX learning procedure (Becker and Hinton, 1992), was also able to learn depth from random dot stereograms by applying a temporal coherence assumption instead of the spatial coherence model described earlier in this chapter. Instead of maximizing mutual information between spatially adjacent outputs, the algorithm maximized the mutual information in a neuron's output at nearby points in time. In a related model, Stone (Stone, 1996) demonstrated that an algorithm that minimized the short term variance of a neuron's output while maximizing its variance over longer time scales also learned to estimate depth in moving random dot stereograms. This algorithm can be shown to be equivalent to IMAX, with more straightfor-

ward implementation (Stone, personal communication). The two algorithms make the assumption that properties of the visual world such as depth vary slowly in time. Stone (Stone, 1996) tested this hypothesis with natural images, and found that although natural images contain sharp depth boundaries at object edges, depth varies slowly the vast majority of the time, and his learning algorithm was able to learn depth estimation from natural graylevel images.

Weinshall and Edelman (Weinshall and Edelman, 1991) applied the assumption of temporal persistence of objects to learn object representations that were invariant to rotations in depth. They first trained a 2 layer network to store individual views of wire-framed objects. Then they updated lateral connections in the output layer with Hebbian learning as the input object rotated through different views. The strength of the association in the lateral connections was proportional to the estimated strength of the perceived apparent motion if the 2 views were presented in succession to a human subject. After training the lateral connections, when one view of an object was presented, the output activity could be iterated until all of the units for that object were active. This formed an attractor network in which each object was associated with a distinct fixed point.[13] When views were presented that differed from the training views, correlation in output ensemble activity decreased linearly as a function of rotation angle from the trained view. This mimicked the linear increase in human response times with rotation away from the memorized view which has been taken has evidence for mental rotation of an internal 3-D object model (Shepard and Cooper, 1982). This provided an existence proof that such responses can be obtained in a system that stores multiple 2-D views. The human data does not prove the existence of internal 3-D object models.

Weinshall and Edelman modeled the development of viewpoint invariance using idealized objects consisting of paper-clip style figures with labeled vertex locations. The temporal coherence assumption has more recently been applied to learning viewpoint invariant representations of objects in graylevel images (Bartlett and Sejnowski, 1996b; Bartlett and Sejnowski, 1997; Wallis and Rolls, 1997; Becker, 1999). Földíak's learning scheme can be applied in a multi-layer multi-resolution network to learn transformation invariant letter recognition (Wallis and Baddeley, 1997), and face recognition that is invariant to rotations in the plane (Wallis and Rolls, 1997). Becker (Becker, 1999) extended a competitive mixture-of-Gaussians learning model (Nowlan, 1990) to include modulation by temporal context. In one simulation, the algorithm learned responses to facial identity independent of viewpoint, and by altering the architecture, a second simulation learned responses to viewpoint independent of

[13] An attractor network is set of interconnected units which exhibits sustained patterns of activity. The simplest form of attractor network contains "fixed points", which are stable activity rates for all units. The range of input patterns that can settle into a given fixed point is its "basin of attraction."

identity. Chapter 7 of this book (Bartlett and Sejnowski, 1997) examines the development of representations of faces that are tolerant to rotations in depth in both a feedforward system based on Földák's learning mechanism, and in a recurrent system related to Weinshall and Edelman's work, in which lateral interconnections formed an attractor network.

4.2. Temporal association in psychophysics and biology

Such models challenge theories that 3-dimensional object recognition requires the construction of explicit internal 3-dimensional models of the object. The models presented by Földíak, Weinshall, O'Reilly & Johnson, and Becker, in which individual output units acquire transformation tolerant representations, suggest another possibility. Representations may consist of several views that contain a high degree of rotation tolerance about a preferred view. It has been proposed that recognition of novel views may instead be accomplished by linear (Ullman and Basri, 1991) or nonlinear combinations of stored 2-D views (Poggio and Edelman, 1990; Bulthoff et al., 1995). Such view-based representations may be particularly relevant for face processing, given the recent psychophysical evidence for face representations based on low-level filter outputs (Biederman, 1998; Bruce, 1998). Face cells in the primate inferior temporal lobe have been reported with broad pose tuning on the order of $\pm 40^{\circ}$ (Perrett et al., 1989; Hasselmo et al., 1989). Perrett and colleagues (Perrett et al., 1989), for example, reported broad coding for five principal views of the head: Frontal, left profile, right profile, looking up, and looking down.

There are several biological mechanisms by which receptive fields could be modified to perform temporal associations. A temporal window for Hebbian learning could be provided by the 0.5 second open-time of the NMDA channel (Rhodes, 1992; Rolls, 1992). A spatio-temporal window for Hebbian learning could also be produced by the release of a chemical signal following activity such as nitric oxide (Montague et al., 1991). Reciprocal connections between cortical regions (O'Reilly and Johnson, 1994) or lateral interconnections within cortical regions could sustain activity over longer time periods and allow temporal associations across larger time scales.

Temporal association may be an important factor in the development of viewpoint invariant responses in the inferior temporal lobe[14] of primates (Rolls, 1995). Neurons in the anterior inferior temporal lobe are capable of forming temporal associations in their sustained activity patterns. After prolonged exposure to a sequence of randomly generated fractal patterns, correlations emerged in the sustained responses to neighboring patterns in the sequence (Miyashita, 1988). These data suggest that cells in the temporal lobe modify

[14]The inferior temporal lobe of primates has been associated with visual object processing and pattern recognition.

their receptive fields to associate patterns that occurred close together in time. This is a mechanism by which cortical neurons could associate different views of an object without requiring explicit three-dimensional representations or complex geometrical transformations (Stryker, 1991b).

Dynamic information appears to play a role in representation and recognition of faces and objects by humans. Human subjects were better able to recognize famous faces when the faces were presented in video sequences, as compared to an array of static views (Lander and Bruce, 1997). Recognition of novel views of unfamiliar faces was superior when the faces were presented in continuous motion during learning (Pike et al., 1997). Stone (Stone, 1998) obtained evidence that dynamic signals contribute to object representations beyond providing structure-from-motion. Recognition rates for rotating amoeboid objects decreased, and reaction times increased when the temporal order of the image sequence was reversed in testing relative to the order during learning.

5. COMPUTATIONAL ALGORITHMS FOR RECOGNIZING FACES IN IMAGES

One of the earliest approaches to recognizing facial identity in images was based on a set of feature measurements such as nose length, chin shape, and distance between the eyes (Kanade, 1977; Brunelli and Poggio, 1993). An advantage of a feature-based approach to image analysis is that it drastically reduces the number of input dimensions, and human intervention can be employed to decide what information in the image is relevant to the task. A disadvantage is that the specific image features relevant to the classification may not be known in advance, and vital information may be lost when compressing the image into a limited set of features. Moreover, holistic graylevel information appears to play an important role on human face processing (Bruce, 1988), and may contain useful information for computer face processing as well. An alternative to feature-based image analysis emphasizes preserving the original images as much as possible and allowing the classifier to discover the relevant features in the images. Such approaches include template matching. Templates capture information about configuration and shape that can be difficult to parameterize. In some direct comparisons of face recognition using feature-based and template-based representations, the template approaches outperformed the feature-based systems (Brunelli and Poggio, 1993; Lanitis et al., 1997). Accurate alignment of the faces is critical to the success of template-based approaches. Aligning the face, however, can be more straightforward than precise localization of individual facial landmarks for feature-based representations.

A variant of the template matching approach is an adaptive approach to image analysis in which image features relevant to facial actions are learned directly from example image sequences. In such approaches to image analysis, the physical properties relevant to the classification need not be specified in

advance, and are learned from the statistics of the image set. This is particularly useful when the specific features relevant to the classification are unknown (Valentin et al., 1994).

An adaptive approach to face image analysis that has achieved success for face recognition is based on principal component analysis of the image pixels (Millward and O'Toole, 1986; Cottrell and Fleming, 1990; Turk and Pentland, 1991). As discussed in Section 1.5, PCA is a form of unsupervised learning related to Hebbian learning that extracts image features from the second order dependencies among the image pixels. PCA is performed on the images by considering each image as a high dimensional observation vector, with the graylevel of each pixel as the measure. The principal component axes are the eigenvectors of the pixelwise covariance matrix of the dataset. These component axes are template images that can resemble ghost-like faces which have been labeled "holons" (Cottrell and Fleming, 1990) and "eigenfaces" (Turk and Pentland, 1991). A low-dimensional representation of the face images with minimum reconstruction error is obtained by projecting the images onto the first few principal component axes, corresponding to the axes with the highest eigenvalues. The projection coefficients constitute a feature vector for classification. Representations based on principal component analysis have been applied successfully to recognizing facial identity (Cottrell and Fleming, 1990; Turk and Pentland, 1991), facial expressions (Cottrell and Metcalfe, 1991; Bartlett et al., 1996; Padgett and Cottrell, 1997), and to classifying the gender of the face (Golomb et al., 1991).

Compression networks, consisting of a three layer network trained to reconstruct the input in the output after forcing the data through a low dimensional "bottleneck" in the hidden layer, perform principal component analysis of the data (Cottrell and Fleming, 1990). The networks are trained by backpropagation to reconstruct the input in the output with minimum squared error. When the transfer function is linear, the N hidden unit activations span the space of the first N principal components of the data. New views of a face can be synthesized from a sample view using principal component representations of face shape and texture. Vetter and Poggio (Vetter and Poggio, 1997) performed PCA separately on the frontal and profile views of a set of face images. Assuming rigid rotation and orthographic projection, they showed that the coefficients for the component axes of the frontal view could be linearly predicted from the coefficients of the profile view axes.

The principal component axes that account for the most reconstruction error, however, are not necessarily the ones that provide the most information for recognizing facial identity. O'Toole and colleagues (O'Toole et al., 1993) demonstrated that the first few principal component axes, which contained low spatial frequency information, were most discriminative for classifying gender, whereas a middle range of components, containing a middle range of spatial

frequencies, were the most discriminative for classifying facial identity. This result is consistent with recordings of the responses of face cells to band-pass filtered face images (Rolls et al., 1987). The face cells in the superior temporal sulcus responded most strongly to face images containing energy in a middle range of spatial frequencies, between 4 and 32 cycles per image.

Principal component analysis is a form of autoassociative memory (Valentin et al., 1994). The PCA network reproduces the input in the output with minimum squared error. Kohonen (Kohonen et al., 1981) was the first to use an autoassociative memory to store and recall face images. Kohonen generated an autoassociative memory for 100 face images by employing a simple Hebbian learning rule. Noisy or incomplete images were then presented to the network, and the images reconstructed by the network were similar in appearance to the original, noiseless images. The reconstruction accuracy of the network can be explicitly measured by the cosine of the angle between the network output and the original face image (Millward and O'Toole, 1986). Reconstructing the faces from an autoassociative memory is akin to applying a Wiener filter to the face images, where the properties of the filter are determined by the "face history" of the weight matrix (Valentin et al., 1994).

In such autoassociative networks, a whole face can be recovered from a partial input, thereby acting as content-addressable memory. Cottrell (Cottrell, 1990) removed a strip of a face image, consisting of about 20% of the total pixels. The principal component-based network reconstructed the face image, and filled in the missing pixels to create a recognizable face. Autoassociative networks also provide a means of handling occlusions. If a PCA network is trained only on face images, and then the presented with a face image that contains an occluding object, such as a hand in front of the face, the network will reconstruct the face image without the occluding object (Cottrell, personal communication). This occurs because the network reconstruction is essentially a linear combination of the images on which the network was trained – the PCA eigenvectors are linear combinations of the original data. Since the occluding object is distant from the portion of image space spanned by the principal component axes, the projection of the face image onto the component axes will be dominated by the face portions of the image, and will reconstruct an image that is similar to the original face. Because the network had no experience with hands, it would be unable to reproduce anything about the hand.

Autoassociative memory in principal component-based networks provides an account for some aspects of human face perception. Principal component representations of face images have been shown to account well for human perception of distinctiveness and recognizability (O'Toole et al., 1994) (Hancock et al., 1996). Such representations have also demonstrated phenomena such as the "other race effect" (O'Toole et al., 1994). Principal component axes trained on a set of faces from one race are less able to capture the directions

of variability necessary to discriminate faces from another race. Eric Cooper has shown that alteration of the aspect ratio of a face interferes strongly with recognition, although the image still looks like a face, whereas displacement of one eye appears significantly distorted, yet interferes only slightly with recognition of the face (Cooper, 1998). A similar effect would be observed in principal component-based representations (Gary Cottrell, personal communication). The elongated face image would still lie within face space; Its distance to the PCA axes would be short, and therefore would be classed as a face. The aspect ratio manipulation, however, would alter the projection coefficients, which would therefore interfere with recognition. Displacement of one eye would cause the image to lie farther from face space, but would have a much smaller effect on the projection coefficients of the face image.

Another holistic spatial representation is obtained by a class-specific linear projection of the image pixels (Belhumeur et al., 1997). This approach is based on Fisher's linear discriminants, which is a supervised learning procedure that projects the images into a subspace in which the classes are maximally separated. A class may be constituted, for example, of multiple images of a given individual under different lighting conditions. Fisher's Linear Discriminant is a projection into a subspace that maximizes the between–class scatter while minimizing the within–class scatter of the projected data. This approach assumes linear separability of the classes. It can be shown that face images under changes in lighting lie in an approximately linear subspace of the image space if we assume the face is modeled by a Lambertian surface (Shashua, 1992; Hallinan, 1995). Fisher's linear discriminant analysis performed well for recognizing faces under changes in lighting. The linear assumption breaks down for dramatic changes in lighting that strongly violate the Lambertian assumption by, for example, producing shadows on the face from the nose. Another limitation of this approach is that projection of the data onto a very few dimensions can make linear separability of test data difficult.

Penev and Atick (Penev and Atick, 1996) developed a topographic representation based on principal component analysis, which they termed "local feature analysis." The representation is based on a set of kernels that are matched to the second-order statistics of the input ensemble. The kernels were obtained by performing a decorrelating "retinal" transfer function on the principal components. This transfer function whitened the principal components, meaning that it equalized the power over all frequencies. The whitening process was followed by a rotation to topographic correspondence with pixel location. An alternative description of the LFA representation is that it is the principal component reconstruction of the image using whitened PCA coefficients. Both the eigenface approach and LFA separate only the second order moments of the images, but do not address the high-order statistics. These image statistics include relationships between three or more pixels, such as edges, curvature, and

shape. In a task such as face recognition, much of the important information may be contained in such high-order image properties.

Classification of local feature measurements is heavily dependent on exactly which features were measured. Padgett & Cottrell (Padgett and Cottrell, 1997) found that an "eigenfeature" representation of face images, based in the principal components of image regions containing individual facial features such as an eye or a mouth, outperformed the full eigenface representation for classifying facial expressions. Best performance was obtained using a representation based on image analysis over even smaller regions. The representation was derived from a set of local basis functions obtained from principal component analysis of subimage patches selected from random image locations. This finding is supported by Gray, Movellan & Sejnowski (Gray et al., 1997) who also obtained better performance for visual speechreading using representations derived from local basis functions.

Another local representation that has achieved success for face recognition is based on the outputs of a banks of Gabor filters. Gabor filters, obtained by convolving a 2-D sine wave with a Gaussian envelope, are local filters that resemble the responses of visual cortical cells (Daugman, 1988). Representations based on the outputs of these filters at multiple spatial scales, orientations, and spatial locations, have been shown to be effective for recognizing facial identity (Lades et al., 1993). Relationships have been demonstrated between Gabor filters and statistical independence. Bell & Sejnowski (Bell and Sejnowski, 1997) found that the filters that produced independent outputs from natural scenes were spatially local, oriented edge filters, similar to a bank of Gabor filters. It has also been shown that Gabor filter outputs of natural images are independent under certain conditions(Simoncelli, 1997).

The elastic matching algorithm (Lades et al., 1993) represents faces usong banks of Gabor filters. It includes a dynamic recognition process that provides tolerance to small shifts in spatial position of the image features due to small changes in pose or facial expression. In a direct comparison of face recognition algorithms, the elastic matching algorithm based on the outputs of Gabor filters gave better face recognition performance than the eigenface algorithm based on principal component analysis (Zhang et al., 1997; Phillips et al., 1998).

The elastic matching paradigm represents faces as a labeled graph, in which each vertex of a 5 x 7 graph stores a feature vector derived from a set of local spatial filters. The filter bank consists of wavelets based on Gabor functions, and covers five spatial frequencies and eight orientations. These feature vectors represent the local power spectrum in the image. The edges of the graph are labeled with the distance vectors between the vertices.

During the dynamic recognition process, all face models in the database are distorted to fit the new input as closely as possible. The vertices of each graph model are positioned at coordinates which maximize the correlation between

the model and the input image, while minimizing the deviation from the original shape of the graph. This elastic match is carried out by optimizing the following cost function, H, for each model M, over positions i in the input image I:

$$H^M(i^I) = \frac{a}{2}\sum_{i,j} D_l(L^I_{ij}, L^M_{ij}) - \sum_i S_v(J^I_i, J^M_i) \tag{2.20}$$

$$\text{where} \qquad D_l(L^I_{ij}, L^M_{ij}) = (L^I_{ij} - L^M_{ij})^2$$
$$S_v(J^I_i, J^M_i) = \frac{J^I_i \cdot J^M_i}{||J^I_i||||J^M_i||}$$

In this cost function, S_v measures the similarity between the feature vector of the model and that of the input image at vertex location i, and D_l is distortion expressed as the squared length of the difference vector between the expected edge vector in the model and the corresponding edge label in the distorted graph. The face model with the best fit is accepted as a match.

The elastic matching paradigm addresses the problem of face alignment and feature detection in two ways. The amplitude of the Gabor filter outputs changes smoothly with shifts in spatial position, so that alignment offsets do not have a catastrophic effect on recognition. Secondly, the elastic matching phase of the algorithm explicitly minimizes the effect of small changes in spatial position of the facial features between the model and the input image by allowing distortions in the node positions.

Chapter 3 introduces face representations based on independent component analysis. Whereas the eigenface and LFA representations learn the second-order dependencies in the image ensemble, the ICA represenation learns the high-order dependencies as well. Gabor wavelets, PCA, and ICA each provide a way to represent face images as a linear superposition of basis functions. PCA models the data as a multivariate Gaussian, and the basis functions are restricted to be orthogonal (Lewicki and Olshausen, 1998). ICA allows the learning of non-orthogonal bases and allows the data to be modeled with non-Gaussian distributions (Comon, 1994). As noted in Section 2.3, there are relationships between Gabor wavelets and the basis functions obtained with ICA (Bell and Sejnowski, 1997). The Gabor wavelets are not specialized to the particular data ensemble, but would be advantageous when the number of data samples is small. The following chapters compare these face analysis algorithms, and addresses issues of hand engineered features versus adaptive features, local vs global spatial analysis, and learning second-order versus all-order dependencies in face images.