Face Image Analysis by Unsupervised Learning

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

Acknowledgments xi

1. SUMMARY 1

2. INTRODUCTION 5
   - 2.1 Unsupervised learning in object representations 5
     - 2.1.1 Generative models 6
     - 2.1.2 Redundancy reduction as an organizational principle 8
     - 2.1.3 Information theory 9
     - 2.1.4 Redundancy reduction in the visual system 11
     - 2.1.5 Principal component analysis 12
     - 2.1.6 Hebbian learning 13
     - 2.1.7 Explicit discovery of statistical dependencies 15
   - 2.2 Independent component analysis 17
     - 2.2.1 Decorrelation versus independence 17
     - 2.2.2 Information maximization learning rule 18
     - 2.2.3 Relation of sparse coding to independence 22
   - 2.3 Unsupervised learning in visual development 24
     - 2.3.1 Learning input dependencies: Biological evidence 24
     - 2.3.2 Models of receptive field development based on correlation sensitive learning mechanisms 26
   - 2.4 Learning invariances from temporal dependencies in the input 29
     - 2.4.1 Computational models 29
     - 2.4.2 Temporal association in psychophysics and biology 32
   - 2.5 Computational Algorithms for Recognizing Faces in Images 33

3. INDEPENDENT COMPONENT REPRESENTATIONS FOR FACE RECOGNITION 39
   - 3.1 Introduction 39
     - 3.1.1 Independent component analysis (ICA) 42
     - 3.1.2 Image data 44
   - 3.2 Statistically independent basis images 45
     - 3.2.1 Image representation: Architecture 1 45
     - 3.2.2 Implementation: Architecture 1 46

3.2.3 Results: Architecture 1 48
3.3 A factorial face code 53
3.3.1 Independence in face space versus pixel space 53
3.3.2 Image representation: Architecture 2 54
3.3.3 Implementation: Architecture 2 56
3.3.4 Results: Architecture 2 56
3.4 Examination of the ICA Representations 59
3.4.1 Mutual information 59
3.4.2 Sparseness 60
3.5 Combined ICA recognition system 62
3.6 Discussion 63

4. AUTOMATED FACIAL EXPRESSION ANALYSIS 69
4.1 Review of other systems 70
4.1.1 Motion-based approaches 70
4.1.2 Feature-based approaches 71
4.1.3 Model-based techniques 72
4.1.4 Holistic analysis 73
4.2 What is needed 74
4.3 The Facial Action Coding System (FACS) 75
4.4 Detection of deceit 78
4.5 Overview of approach 81

5. IMAGE REPRESENTATIONS FOR FACIAL EXPRESSION ANALYSIS: COMPARATIVE STUDY I 83
5.1 Image database 84
5.2 Image analysis methods 85
5.2.1 Holistic spatial analysis 85
5.2.2 Feature measurement 87
5.2.3 Optic flow 88
5.2.4 Human subjects 90
5.3 Results 91
5.3.1 Hybrid system 93
5.3.2 Error analysis 94
5.4 Discussion 96

6. IMAGE REPRESENTATIONS FOR FACIAL EXPRESSION ANALYSIS: COMPARATIVE STUDY II 101
6.1 Introduction 102
6.2 Image database 103
6.3 Optic flow analysis 105
6.3.1 Local velocity extraction 105
6.3.2 Local smoothing 105
6.3.3 Classification procedure 106
6.4 Holistic analysis 108
6.4.1 Principal component analysis: “EigenActions” 108
6.4.2 Local feature analysis (LFA) 109

6.4.3 "FisherActions" 112
6.4.4 Independent component analysis 114
6.5 Local representations 117
6.5.1 Local PCA 117
6.5.2 Gabor wavelet representation 119
6.5.3 PCA jets 120
6.6 Human subjects 122
6.7 Discussion 123
6.8 Conclusions 127

7. LEARNING VIEWPOINT INVARIANT REPRESENTATIONS OF FACES 129
7.1 Introduction 129
7.2 Simulation 133
7.2.1 Model architecture 134
7.2.2 Competitive Hebbian learning of temporal relations 134
7.2.3 Temporal association in an attractor network 137
7.2.4 Simulation results 140
7.3 Discussion 147

8. CONCLUSIONS AND FUTURE DIRECTIONS 151

References 157

Index 169

This book is dedicated to
Nigel.

Acknowledgments

This book evolved from my doctoral dissertation at the University of California, San Diego. It was a great privilege to work with my thesis adviser, Terry Sejnowski, for five years at the Salk Institute. I benefited enormously from his breadth of knowledge and capacity for insight, and from the diverse and energetic laboratory environment that he created at the Salk Institute. An important thanks goes to my Committee Chair, Don Macleod, for his encouragement throughout this interdisciplinary thesis. With his remarkable breadth of knowledge, he provided invaluable advice and guidance at many important points in my graduate education. I would also like to thank Javier Movellan for encouraging me to write this book, and for providing a motivating research environment at UCSD in which to pursue the next phases of this research. I am grateful to Gary Cottrell for giving a tremendous Cognitive Science lecture series on face recognition which provided the foundation for much of the work that appears in this book. I am also endebted to Gary for referring my thesis to Kluwer. This book would not have materialized without him. Most of the research presented in Chapter 6 was conducted by my colleague, Gianluca Donato. It was a privilege to work with such a productive and congenial researcher. I also thank my office-mate Michael Gray for sharing ideas, space, and experiences over more than five years of graduate school. I am grateful to my parents, whose limitless supply of support and encouragement sustained me throughout my thesis work. My biggest debt of gratitude goes to Nigel for his love and support throughout this endeavor, and to our son, Paul, for keeping things in perspective by bouncing in his jumping chair while I was writing.

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

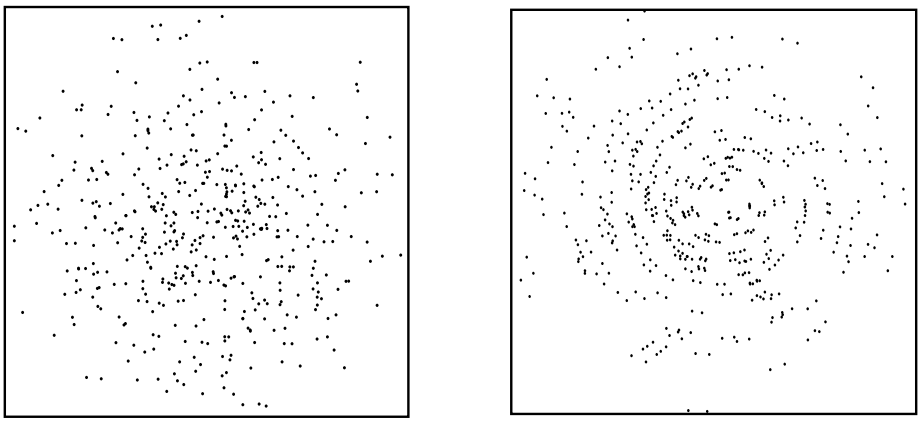

Figure 2.1. The percept of structure is driven by the dependencies. LEFT: A set of points selected from a Gaussian distribution. RIGHT: Half of the points were selected from a Gaussian distribution, and the other half were generated by rotating the points 5° about the centroid of the distribution. Figure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

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

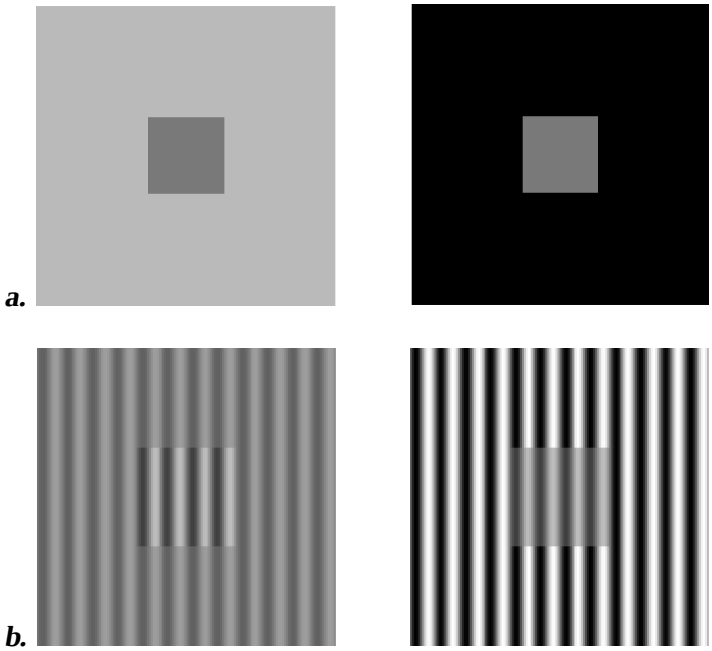

Figure 2.2. Redundancy reduction in the visual system. a. Luminance adaptation. The center squares are the same shade of gray, but the square on the left appears darker than the square on the right. b. Contrast adaptation. The center squares have the same contrast, but the square on the left appears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

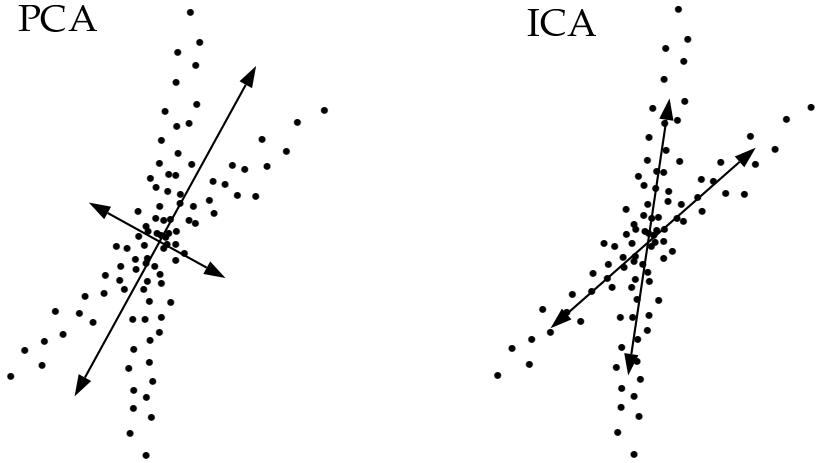

Figure 2.3. Example 2-D data distribution and the corresponding principal component and independent component axes. The data points could be, for example, grayvalues at pixel 1 and pixel 2. Figure inspired by Lewicki & Sejnowski (2000).

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

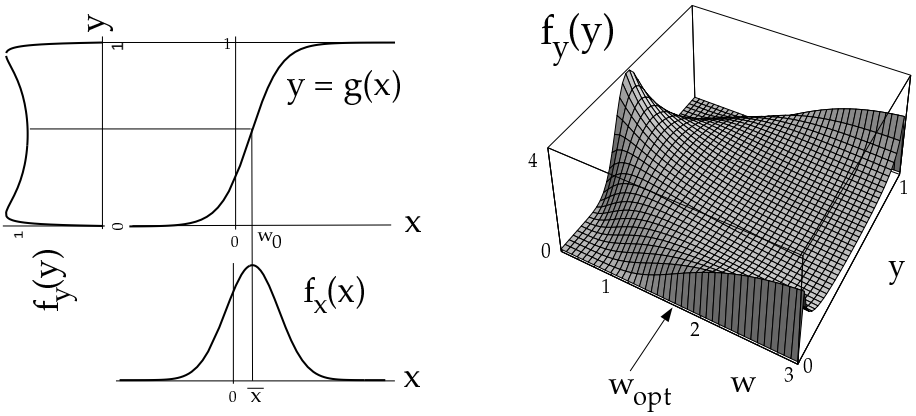

Figure 2.4. Optimal information flow in sigmoidal neurons. The input x is passed through a nonlinear function, $g(x)$. The information in the output density $f_y(y)$ depends on matching the mean and variance of $f_x(x)$ to the slope and threshold of $g(x)$. Right: $f_y(y)$ is plotted for different values 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.

Chapter 3

INDEPENDENT COMPONENT REPRESENTATIONS FOR FACE RECOGNITION

Abstract In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels. A number of face recognition algorithms employ principal component analysis (PCA), which is based on the second-order statistics of the image set, and does not address high-order statistical dependencies such as the relationships among three or more pixels. Independent component analysis (ICA) is a generalization of PCA which separates the high-order moments of the input in addition to the second-order moments. ICA was performed on a set of face images by an unsupervised learning algorithm derived from the principle of optimal information transfer through sigmoidal neurons (Bell and Sejnowski, 1995). The algorithm maximizes the mutual information between the input and the output, which produces statistically independent outputs under certain conditions. ICA was performed on the face images under two different architectures, one which separated images across spatial location, and a second which separated the feature code across images. The first architecture provided a statistically independent basis set for the face images that can be viewed as a set of independent facial feature images. The second architecture provided a factorial code, in which the probability of any combination of features can be obtained from the product of their individual probabilities. Both ICA representations were superior to representations based on principal components analysis for recognizing faces across days and changes in expression.

1. INTRODUCTION

Horace Barlow has argued that redundancy provides knowledge (Barlow, 1989). Redundancy in the sensory input contains structural information about the environment. What is important for the perceptual system to detect is "suspicious coincidences," new statistical regularities in the sensory input that differ from the environment to which it has been adapted. Bars and edges, for example, are locations in the visual input at which there is phase alignment across multiple spatial scales, and constitute a "suspicious coincidence" in Barlow's

formulation (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 these new regularities. 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. Learning such a transformation is equivalent to modeling the prior knowledge of the statistical dependencies in the input. 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.

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 the “simultaneous contrast” illusion, and by contrast gain control mechanisms observed 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 compact, decorrelated representations as a general coding strategy for the visual system.

Some of the most successful algorithms for face recognition, such as "eigenfaces" (Turk and Pentland, 1991), "holons" (Cottrell and Metcalfe, 1991), and “local feature analysis” (Penev and Atick, 1996) are based on learning mechanisms that are sensitive to the correlations in the face images. These are data-driven representations based on principal component analysis of the image set. Principal component analysis removes the correlations in the input, but does not address the high-order dependencies the images, such as the relationships among three or more pixels. Edges are an example of a high-order dependency in an image, as are elements of shape and curvature. In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels.

Second-order statistics capture the amplitude spectrum of images but not the phase. The high order statistics correspond to the phase spectrum (Field, 1994; Bell and Sejnowski, 1997). 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). 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 (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. Bell and Sejnowski (Bell and Sejnowski, 1995; Bell and Sejnowski, 1997) recently developed an algorithm for separating the statistically independent components of a dataset through information maximization. This algorithm has proven successful for separating randomly mixed auditory signals (the cocktail party problem), and has recently been applied to separating EEG signals (Makeig et al., 1996), fMRI signals (McKeown et al., 1998).

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).

This chapter presents methods for representing face images for face recognition based on the statistically independent components of the image set. We performed independent component analysis on the image set under two architectures. The first architecture separated images across space (pixel location). This found a set of independent source images, or facial feature images, in which the pixel values of one feature image cannot be predicted from the pixel values of the other feature images. These source images comprised an independent basis set for the faces, where each face image can be decomposed into a linear superposition of independent source images. The face representation consisted of the coefficients for the linear combination of independent basis images that comprised each face image. This architecture corresponded to the one used to perform blind separation of a mixture auditory signals (Bell and Sejnowski, 1995) and to examine the independent sources of EEG (Makeig et al., 1996) and fMRI data (McKeown et al., 1998). Under this architecture, the basis images were independent, but the coding variables that represented each face image were not. The second architecture found a factorial face code. It defined a set of statistically independent coding variables for representing the face images. This architecture separated pixels across images, and corresponded to the architecture used to find image filters that produced statistically independent outputs from natural scenes (Bell and Sejnowski, 1997). Such a factorial code can be advantageous for encoding complex objects that are characterized by high-order combinations of features, since the prior probability of any combination of features can be obtained from their individual probabilities (Barlow, 1989; Atick, 1992). Mat-

lab code for the ICA representations is available at http://ergo.ucsd.edu/~marni and http://www.cnl.salk.edu/~tewon/ica_cnl.html.

Face recognition performance was tested using the FERET database (Phillips et al., 1998). Face recognition performances using the ICA representations were benchmarked by comparing them to performances using principal component analysis, which is equivalent to the "eigenfaces" representation (Turk and Pentland, 1991; Pentland et al., 1994).

1.1. Independent component analysis (ICA)

Bell and Sejnowski's ICA algorithm is an unsupervised learning rule that was derived from the principle of optimal information transfer through sigmoidal neurons (Laughlin, 1981; Bell and Sejnowski, 1995). This algorithm addresses the case of an arbitrary input, x, and output, y, passed through a nonlinear squashing function, g.

$$u = wx + w_0 \qquad y = g(u) = \frac{1}{1 + e^{-u}} \tag{3.1}$$

The optimal weight w on x for maximizing information transfer is the one that maximizes the entropy of the output. This optimal weight is found by gradient ascent on the entropy of the output, y with respect to w. More information on this learning rule is presented in Chapter 2, Section 2.

When there are multiple inputs and outputs, maximizing the joint entropy of the output encourages the individual outputs to move towards statistical independence. When the form of the nonlinear transfer function g is the same as the cumulative density functions of the underlying independent components (up to scaling and translation) it can be shown that maximizing the mutual information between the input $X = (x_1, x_2, ...)$ and the output $Y = (y_1, y_2, ...)$ also minimizes the mutual information between the u_i (Nadal and Parga, 1994; Bell and Sejnowski, 1997). The update rule for the weight matrix, W, for multiple inputs and outputs is given by

$$\Delta W = (I + y' u^T) W \tag{3.2}$$

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

We employed the logistic transfer function, $g(u) = \frac{1}{1+e^{-u}}$, giving $y' = (1 - 2y_i)$. The logistic transfer function has been found sufficient to separate mixtures of super-Gaussian signals, meaning that the kurtosis of the probability distribution exceeds that of a Gaussian (Bell and Sejnowski, 1995). Many natural signals, such as sound sources, have been shown to have a super-Gaussian distribution.

The algorithm includes a "sphering" step prior to learning (Bell and Sejnowski, 1997). The row means are subtracted from the dataset, X, and then

X is passed through the zero-phase whitening filter, W_z, which is twice the inverse square root of the covariance matrix:

$$W_z = 2 * \langle XX^T \rangle^{-\frac{1}{2}}. \tag{3.3}$$

This removes both the first and the second-order statistics of the data; both the mean and covariances are set to zero and the variances are equalized. The full transform from the zero-mean input was calculated as the product of the sphering matrix and the learned matrix, $W_I = W * W_Z$. The pre-whitening filter in the ICA algorithm has the Mexican-hat shape of retinal ganglion cell receptive fields which remove much of the variability due to lighting (Bell and Sejnowski, 1997).

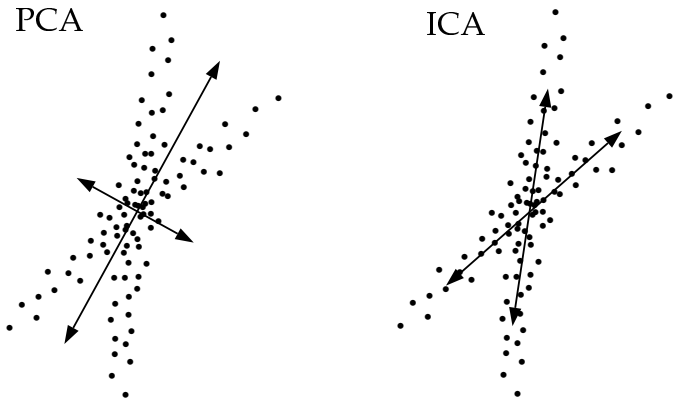

Figure 3.1. Example 2-D data distribution and corresponding principal component and independent component axes. The data points could be, for example, grayvalues at pixel 1 and pixel 2. Figure inspired by Lewicki & Sejnowski (2000).

Some of the differences between PCA and ICA are illustrated as follows. Consider a set of data points derived from two underlying distributions as shown in Figure 3.1. Principal component analysis encodes second order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. PCA constrains the axes to be orthogonal, and in this case, the projections of 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 maximal 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). See Chapter 2, Section 2 for a more detailed comparison of PCA and ICA.

1.2. Image data

The face images employed for this research were a subset of the FERET face database (Phillips et al., 1998). The data set contained images of 425 individuals. There were up to four frontal views of each individual: A neutral expression and a change of expression from one session, and a neutral expression and change of expression from a second session that occurred up to two years after the first. Examples of the four views are shown in Figure 3.2. The algorithms were trained on a single frontal view of each individual. The training set was comprised of 50% neutral expression images and 50% change of expression images. The algorithms were tested for recognition under three different conditions: same session, different expression; different day, same expression; and different day, different expression (see Table 3.1).

Figure 3.2. Example from the FERET database of the four frontal image viewing conditions: Neutral expression and change of expression from Session 1; Neutral expression and change of expression from Session 2. Reprinted with Permission from Jonathon Phillips

Image Set	***Condition***		***No. Images***
Training Set	Session I	50% neutral 50% other	425
Test Set 1	Same Day	Different Expression	421
Test Set 2	Different Day	Same Expression	45
Test Set 3	Different Day	Different Expression	43

Table 3.1. Image sets used for training and testing.

Coordinates for eye and mouth locations were provided with the FERET database. These coordinates were used to center the face images, and then crop and scale them to 60×50 pixels. Scaling was based on the area of the triangle defined by the eyes and mouth. The luminance was normalized by linearly rescaling each image to the interval $[0, 255]$. For the subsequent analyses, the rows of the images were concatenated to produce 1×3000 dimensional vectors. Each image was thus represented as a point in a 3000 dimensional space determined by the luminance value at each pixel location.

2. STATISTICALLY INDEPENDENT BASIS IMAGES

2.1. Image representation: Architecture 1

To find a set of statistically independent basis images for the set of faces, we separated the independent components of the face images according to the image synthesis model of Figure 3.3. The face images in X were assumed to be a linear mixture of an unknown set of statistically independent source images S, where A is an unknown mixing matrix. The sources were recovered by a matrix of learned filters, W_I, which produced statistically independent outputs, U. This synthesis model is related to the one used to perform blind separation on an unknown mixture of auditory signals (Bell and Sejnowski, 1995) and to separate the sources of EEG signals (Makeig et al., 1996) and fMRI images (McKeown et al., 1998).

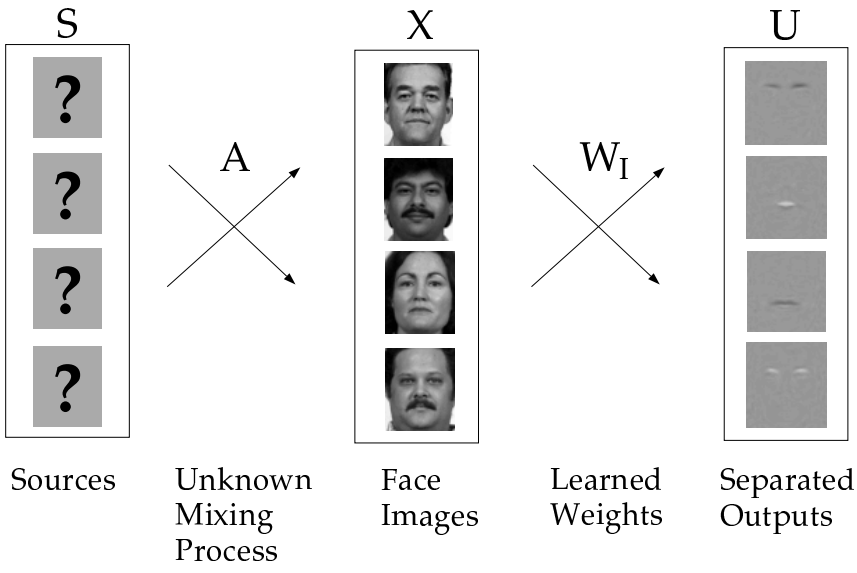

Figure 3.3. Image synthesis model. For finding a set of independent component images, the images in X are considered to be a linear combination of statistically independent basis images, S, where A is an unknown mixing matrix. The basis images were recovered by a matrix of learned filters, W_I, that produced statistically independent outputs, U.

The images comprised the rows of the input matrix, X. With the input images in the rows of X, the ICA outputs in the rows of $W_I X = U$ were also images, and provided a set of independent basis images for the faces (Figure 3.4). These basis images can be considered a set of statistically independent facial features, where the pixel values in each feature image cannot be predicted from the pixel values in the other feature images. The ICA representation consisted of the coefficients for the linear combination of independent basis images in U that

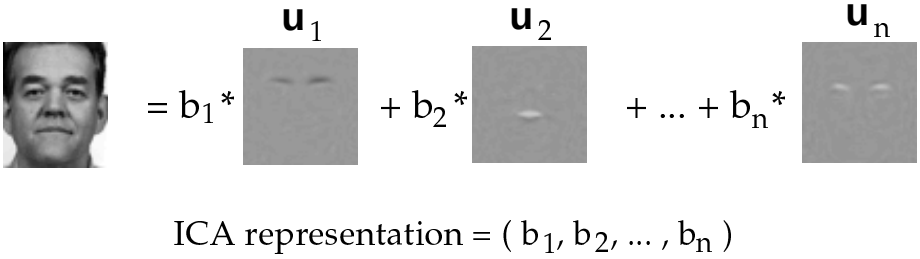

Figure 3.4. The independent basis image representation consisted of the coefficients, $\mathbf{b}$, for the linear combination of independent basis images, $\mathbf{u}$, that comprised each face image $\mathbf{x}$.

comprised each face image, as shown in Figure 3.4. In this model, the matrix of coefficients, B, is obtained from the mixing matrix $A \stackrel{\Delta}{=} W_I^{-1}$. Each row of A contains the coefficients $\mathbf{b}$ for one image $\mathbf{x}$.

2.2. Implementation: Architecture 1

The number of independent components found by the ICA algorithm corresponds to the dimensionality of the input. Since we had 425 image in the training set, the algorithm would attempt to separate 425 independent components. Although we found in previous work that performance improved with the number of components separated, 425 was intractable under our present memory limitations. In order to have control over the number of independent components extracted by the algorithm, instead of performing ICA on the n original images, we performed ICA on a set of m linear combinations of those images, where $m < n$. Recall that the image synthesis model assumes that the images in X are a linear combination of a set of unknown statistically independent sources. The image synthesis model is unaffected by replacing the original images with some other linear combination of the images.

Adopting a method that has been applied to independent component analysis of fMRI data (McKeown et al., 1998), we chose for these linear combinations the first m principal component eigenvectors of the image set. Principal component analysis on the image set in which the pixel locations are treated as observations and each face image a measure, gives the linear combination of the parameters (images) that accounts for the maximum variability in the observations (pixels). The use of PCA vectors in the input did not throw away the high-order relationships. These relationships still existed in the data but were not separated.

Let P_m denote the matrix containing the first m principal component axes in its columns. We performed ICA on P_m^T, producing a matrix of m independent source images in the rows of U. In this implementation, the coefficients, $\mathbf{b}$, for

the linear combination of basis images in U that comprised the face images in X were determined as follows:

The principal component representation of the set of zero-mean images in X based on P_m is defined as $R_m = X * P_m$. A minimum squared error approximation of X is obtained by $X_{rec} = R_m * P_m^T$.

The ICA algorithm produced a matrix $W_I = W * W_Z$ such that

$$W_I * P_m^T = U \qquad \Rightarrow \qquad P_m^T = W_I^{-1} U. \tag{3.4}$$

Therefore

$$X_{rec} = R_m * P_m^T \qquad \Rightarrow \qquad X_{rec} = R_m * W_I^{-1} U. \tag{3.5}$$

where W_Z was the sphering matrix defined in Equation 3.3. Hence the rows of $R_m * W_I^{-1}$ contained the coefficients for the linear combination of statistically independent sources U that comprised X_{rec}, where X_{rec} was a minimum squared error approximation of X, just as in PCA. The independent component representation of the face images based on the set of m statistically independent feature images, U was therefore given by the rows of the matrix

$$B = R_m * W_I^{-1}. \tag{3.6}$$

A representation for test images was obtained by using the principal component representation based on the training images to obtain $R_{test} = X_{test} * P_m$, and then computing $B_{test} = R_{test} * W_I^{-1}$.

Note that the PCA step is not required for the ICA representation of faces. It was employed to serve two purposes: 1. To reduce the number of sources to a tractable number, and 2. To provide a convenient method for calculating representations of test images. Without the PCA step, $B = W_I^{-1}$ and B_{test} can be obtained without calculating a pseudo-inverse by normalizing the length of the rows of U, thereby making U approximately orthonormal[1], and calculating $B_{test} = X_{test} * U^T$.

The principal component axes of the Training Set were found by calculating the eigenvectors of the pixelwise covariance matrix over the set of face images. Independent component analysis was then performed on the first 200 of these eigenvectors, P_{200}, where the first 200 principal components accounted for over 98% of the variance in the images.[2] The 1×3000 eigenvectors in P_{200} comprised the rows of the 200×3000 input matrix X. The input matrix X was sphered according to Equation 3.3, and the weights, W, were updated

[1] A limitation is that if ICA did not remove all of the second-order dependencies then U will not be precisely orthonormal.

[2] In pilot work, we found that face recognition performance improved with the number of components separated. We chose 200 components as the largest number to separate within our processing limitations.

according to Equation 3.2 for 1600 iterations. The learning rate was initialized at 0.001 and annealed down to 0.0001. Training took 90 minutes on a Dec Alpha 2100a. Following training, a set of statistically independent source images were contained in the rows of the output matrix U.

Figure 3.5 shows a subset of 25 source images. A set of principal component basis images (PCA axes), are shown in Figure 3.6 for comparison. The ICA basis images were more spatially local than the principal component basis images. Two factors contribute to the local property of the ICA basis images: The majority of the statistical dependencies were in spatially proximal image locations, and ICA algorithm produces sparse outputs (Bell and Sejnowski, 1997).

These source images in the rows of U were used as the basis of the ICA representation. The coefficients for the zero-mean training images were contained in the rows of $B = R_{200} * W_I^{-1}$ according to Equation 3.6, and coefficients for the test images were contained in the rows of $B_{test} = R_{Test} * W_I^{-1}$ where $R_{Test} = Xtest_i * P_{200}$.

Face recognition performance was evaluated for the coefficient vectors $\mathbf{b}$ by the nearest neighbor algorithm. Coefficient vectors in each test set were assigned the class label of the coefficient vector in the training set that was most similar as evaluated by the cosine of the angle between them:

$$d = \frac{\mathbf{b}_{test} \cdot \mathbf{b}_{train}}{\| \mathbf{b}_{test} \| \| \mathbf{b}_{train} \|}. \tag{3.7}$$

Face recognition performance for the principal component representation was evaluated by an identical procedure, using the principal component coefficients contained in the rows of R_{200}.

2.3. Results: Architecture 1

Figure 3.7 gives face recognition performance with both the ICA and the PCA based representations. Recognition performance is also shown for the PCA based representation using the first 20 principal component vectors, which was the eigenface representation used by Pentland, Moghaddam and Starner (Pentland et al., 1994). Best performance for PCA was obtained using 200 coefficients. Excluding the first 1, 2, or 3 principal components did not improve PCA performance, nor did selecting intermediate ranges of components from 20 through 200. There was a trend for the ICA representation to give superior face recognition performance to the PCA representation with 200 components. The difference in performance was marginally significant for Test Set 3 ($Z = 1.94, p = 0.05$). The difference in performance between the ICA representation and the eigenface representation with 20 components was statistically significant over all three test sets ($Z = 2.5, p < 0.05$) for Test sets 1 and 2, and ($Z = 2.4, p < 0.05$) for Test Set 3.

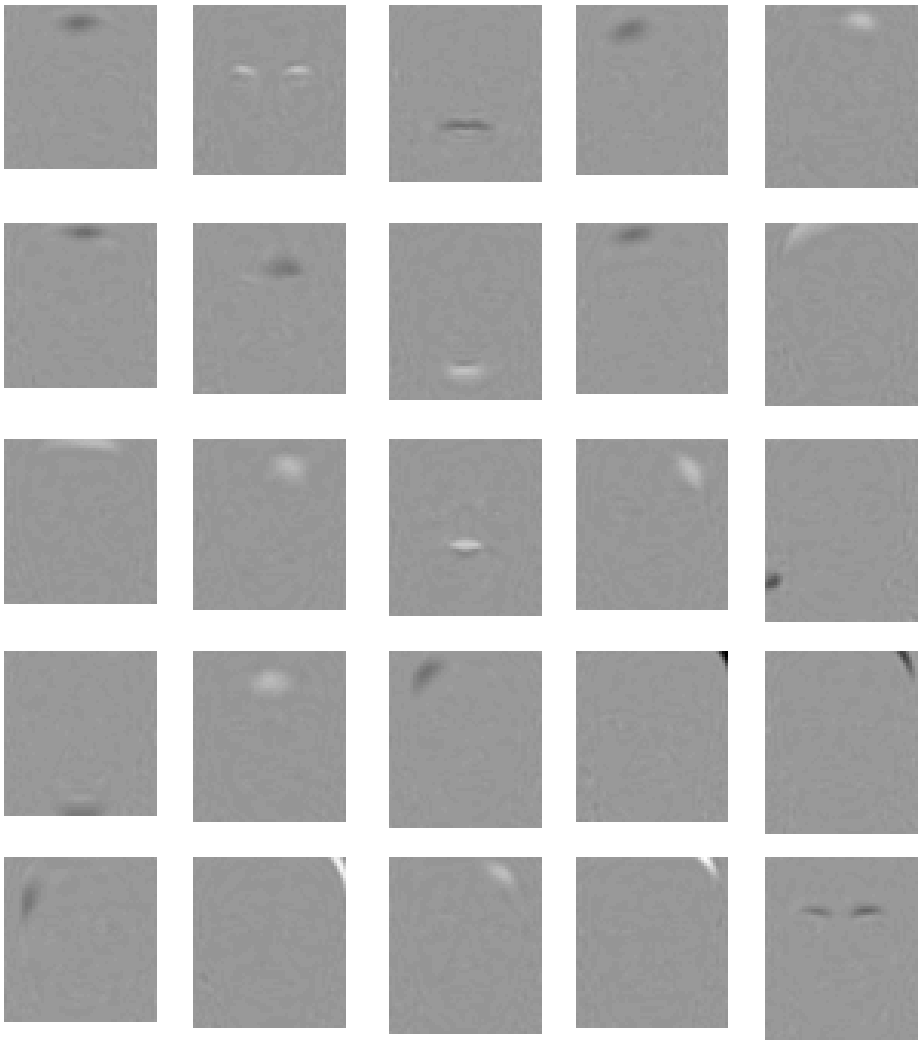

Figure 3.5. Twenty-five independent components of the image set obtained by Architecture 1, which provide a set of statistically independent basis images (rows of U in Figure 3.3). Independent components are ordered by the class discriminability ratio, r (Equation 3.8).

Recognition performance using different numbers of independent components was also examined by performing ICA on 20 to 200 image mixtures in steps of 20. Best performance was obtained by separating 200 independent components, and in general, the more independent components were separated, the better the recognition performance. The basis images also became increasingly spatially local as the number of separated components increased.

Figure 3.6. First 25 principal component axes of the image set (columns of P), ordered left to right, top to bottom, by the magnitude of the corresponding eigenvalue.

Face recognition performances for the PCA and ICA representations were next compared by selecting subsets of the 200 components by class discriminability. Let $\overline{x}$ be the overall mean of a coefficient b_k across all faces, and $\overline{x}_j$ be the mean for person j. For both the PCA and ICA representations, we calculated the ratio of between-class to within-class variability, r, for each coefficient:

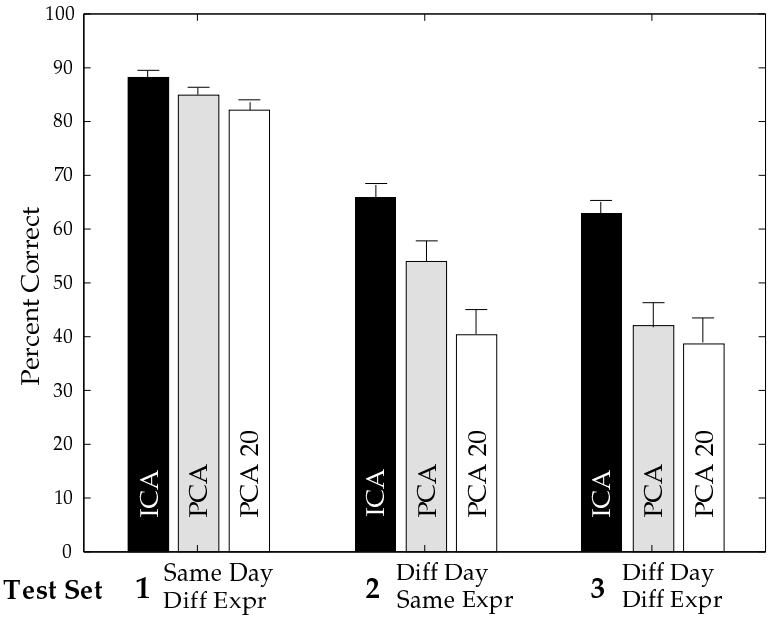

Figure 3.7. Percent correct face recognition for the ICA representation using 200 independent components, the PCA representation using 200 principal components, and the PCA representation using 20 principal components. Groups are performances for Test Set 1, Test Set 2, and Test Set 3. Error bars are one standard deviation of the estimate of the success rate for a Bernoulli distribution.

$$r = \frac{\sigma_{between}}{\sigma_{within}} \tag{3.8}$$

where $\sigma_{between} = \sum_j (\overline{x}_j - \overline{x})^2$ is the variance of the j class means, and $\sigma_{within} = \sum_j \sum_i (x_{ij} - \overline{x}_j)^2$ is the sum of the variances within each class.

The class discriminability analysis was carried out using the 43 subjects for which four frontal view images were available. The ratios r were calculated separately for each test set, excluding the test images from the analysis. Both the PCA and ICA coefficients were then ordered by the magnitude of r. Figure 3.8 (Top) compares the discriminability of the ICA coefficients to the PCA coefficients. The ICA coefficients consistently had greater class discriminability than the PCA coefficients.

Face classification performance was compared using the k most discriminable components of each representation. Figure 3.8 (Bottom) shows the best classification performance obtained for the PCA and ICA representations, which was with the 60 most discriminable components for the ICA representation, and the 140 most discriminable components for the PCA representation.

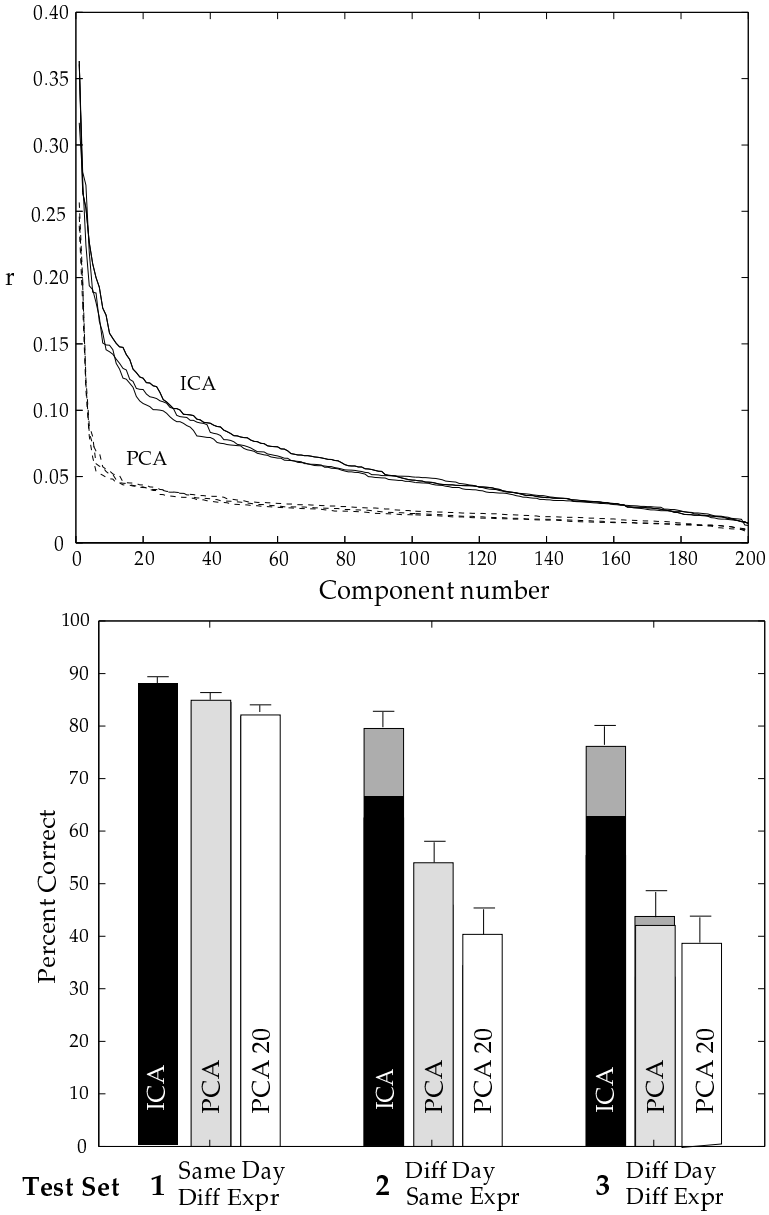

Figure 3.8. Selection of components by class discriminability. Top: Discriminability of the ICA coefficients (solid lines) and discriminability of the PCA components (dotted lines) for the three test cases. Components were sorted by the magnitude of r. Bottom: Improvement in face recognition performance for the ICA and PCA representations using subsets of components selected by the class discriminability r. The improvement is indicated by the gray segments at the top of the bars.

Selecting subsets of coefficients by class discriminability improved the performance of the ICA representation, but had little effect on the performance of the PCA representation. The ICA representation again outperformed the PCA representation. The difference in recognition performance between the ICA and PCA representations was significant for Test Set 2 and Test Set 3, the two conditions that required recognition of images collected on a different day from the training set ($Z = 2.9, p < .05; Z = 3.4, p < .01$), respectively.

3. A FACTORIAL FACE CODE

3.1. Independence in face space versus pixel space

The analysis in Section 2 produced statistically independent basis images. The ICA algorithm separated images across pixel location (see Figure 3.9 Top Left.) Each pixel location was an observation which took on different grayvalues for each of the faces. This is illustrated in Figure 3.9 (Bottom Left), in which the pixels are plotted according to their grayvalues for each face image. ICA in Architecture 1 finds weight vectors in the directions of statistical dependencies in the population of face images over the pixel locations. Projecting the data onto these weights produced a set of independent images, where the pixel grayvalues in one image could not be predicted from the grayvalues of the other images. These independent images spanned the subspace of the face images defined by the first 200 PCA eigenvectors, and each face was represented by the coefficients for the linear combination of these independent template images that comprised each face image.

Although the basis images obtained in Architecture 1 were spatially independent, the coefficients that coded each face were not. By altering the architecture of the independent component analysis, we defined a second representation in which the *coefficients* were statistically independent. In other words, the second ICA architecture found a factorial code for the face images. The alteration in architecture corresponded to transposing the input (see Figure 3.9 Top Right). Each face image was treated as an observation coded by the grayvalues at each of the pixel locations. ICA in Architecture 2 finds weight vectors in the directions of statistical dependencies in the face code across the population of faces. Projecting the data onto these weights produced a set of independent coding variables to replace "pixel location", where the value of any given coding variable could not be predicted from the other coding variables. Each face was represented by the values taken on by this new set of independent coding variables.

The correspondence of the ICA-factorial representation (ICA2) with the principal component representation is direct. The principal component coefficients constitute an uncorrelated face code, whereas the ICA2 coefficients constitute an independent face code.

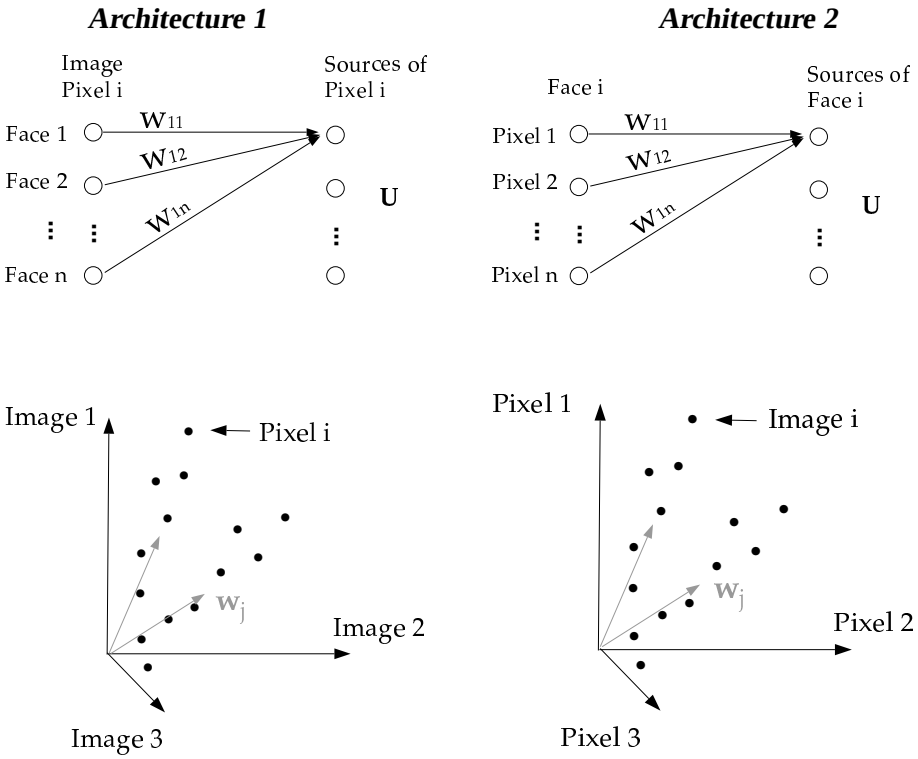

Figure 3.9. Two architectures for performing ICA on images. LEFT: Architecture for finding statistically independent basis images. Top Left: Performing source separation on the face images produced independent component images in the rows of U. Bottom left: The grayvalues at pixel location i are plotted for each face image. ICA in architecture 1 finds weight vectors in the directions of statistical dependencies among the pixel locations. RIGHT: Architecture for finding a factorial code. Top Right: Performing source separation on the pixels produced a factorial code in the columns of the output matrix, U. Bottom Right: Each face image is plotted according to the grayvalues taken on at each pixel location. ICA in architecture 2 finds weight vectors in the directions of statistical dependencies among the face images.

3.2. Image representation: Architecture 2

A factorial face code was obtained by performing source separation on the face images under Architecture 2. The alteration in architecture corresponded to transposing the input matrix X such that the images were in columns and the pixels in rows (see Figure 3.9 Right). Under this architecture, the filters (rows of W_I) were images, as were the columns of $A = W_I^{-1}$. The columns of A formed a new set of basis images for the faces, and the coefficients for reconstructing each face were contained in the columns of the ICA outputs, U.

Architecture 2 is associated with the image synthesis model in Figure 3.11. This model is similar to the the model in Figure 3.3, except that we now

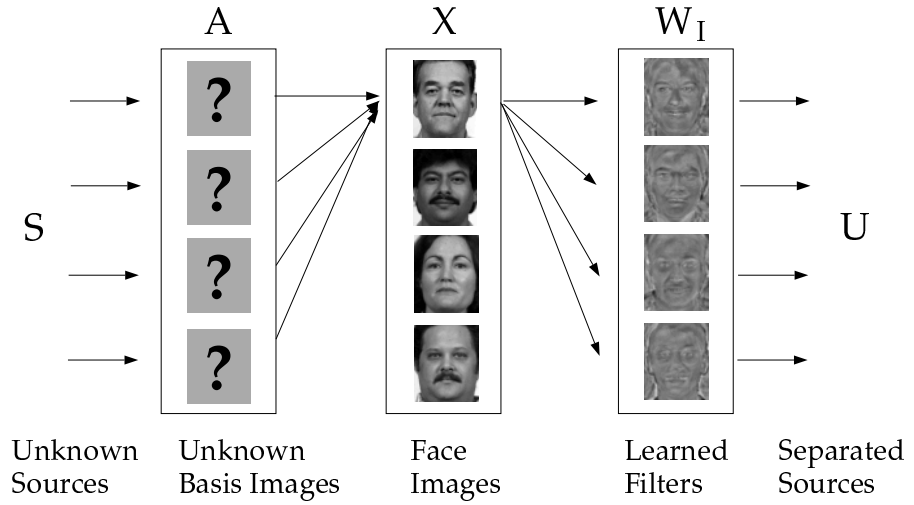

Figure 3.10. Image synthesis model for Architecture 2, based on Olshausen & Field (1996) and Bell & Sejnowski (1997). Each image in the dataset was considered to be a linear combination of underlying basis images in the matrix A. The basis images were each associated with a set of independent "causes", given by a vector of coefficients in S. The causes were recovered by a matrix of learned filters, W_I, which attempts to invert the unknown basis functions to produce statistically independent outputs, U.

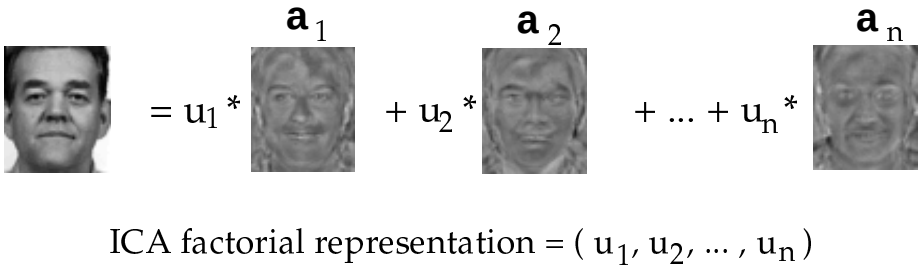

Figure 3.11. The factorial code representation consisted of the independent coefficients, $\mathbf{u}$, for the linear combination of basis images in A that comprised each face image $\mathbf{x}$.

assume that the faces are comprised of a set of independent coefficients, S, for a set of basis images in A, whereas in the model in Figure 3.3 it was the other way around: The independent sources S were basis images, and the coefficients were in A. This model was based on the image synthesis model of Olshausen and Field (Olshausen and Field, 1996b), and was also employed by Bell and Sejnowski (Bell and Sejnowski, 1997) to find image filters that produced statistically independent outputs from natural scenes. The

ICA algorithm attempts to recover the source coefficients by finding a set of filters W_I that produce statistically independent outputs, U.

The columns of the ICA output matrix, $W_I X = U$ provided a factorial code for the training images in X. Each column of U contained the coefficients of the the basis images in A for reconstructing each image in X (Figure 3.11). The representational code for test images was found by $W_I X_{test_i} = U_{test_i}$, where X_{test} was the zero-mean matrix of test images, and W_I was the weight matrix found by performing ICA on the training images.

3.3. Implementation: Architecture 2

ICA was performed on the face images using Architecture 2 to find independent coding variables across images. Placing the pixel values themselves in the columns of X would cause the ICA algorithm to attempt to extract 3000 independent components. Instead of performing ICA directly on the 3000 image pixels, ICA was performed on the first 200 PCA coefficients of the face images in order to reduce the dimensionality. The first 200 principal components accounted for over 98% of the variance in the images. These coefficients comprised the columns of the input data matrix, $X = R_{200}^T$.

The ICA algorithm found a 200 $\times$ 200 weight matrix W_I that produced a set of independent coefficients in the output. The basis functions for this representation consisted of the columns of $A = W_I^{-1}$. A sample of the basis set is shown in Figure 3.12, where the principal component reconstruction $P_{200}A$ was used to visualize the bases as images. The basis images in A have more global properties than the basis images in the ICA output of Architecture 1 (Figure 3.5). Unlike the ICA output, U, the algorithm does not force the columns of A to be either sparse or independent.

The columns of U contained the representational codes for the training images. The representational code for the test images was found by $W_I X_{test} = U_{test}$, where X_{test} was the zero-mean matrix of the test images. This produced 200 coefficients for each face image, consisting of the outputs of the 200 ICA filters.

3.4. Results: Architecture 2

Face recognition performance was again evaluated by the nearest neighbor procedure. Figure 3.13 compares the face recognition performance using the ICA factorial code representation to the independent basis representation of Section 2 and to the PCA representation, each with 200 coefficients. Again, there was a trend for the ICA factorial representation (ICA2) to outperform the PCA representation for recognizing faces across days. The difference in performance for Test Set 2 is significant ($Z = 2.7, p < 0.01$). There was no significant difference in the performances of the two ICA representations.

Figure 3.12. Basis images for the ICA factorial representation (columns of $A = W_I^{-1}$) obtained with Architecture 2. (See Figure 3.10).

Class discriminability of the 200 ICA factorial coefficients was calculated according to Equation 3.8. Unlike the coefficients in the independent basis representation, the ICA factorial coefficients did not differ substantially from each other according to discriminability r. Selection of subsets of components for the representation by class discriminability had little effect on the recognition performance using the ICA-factorial representation (see Figure 3.14). The

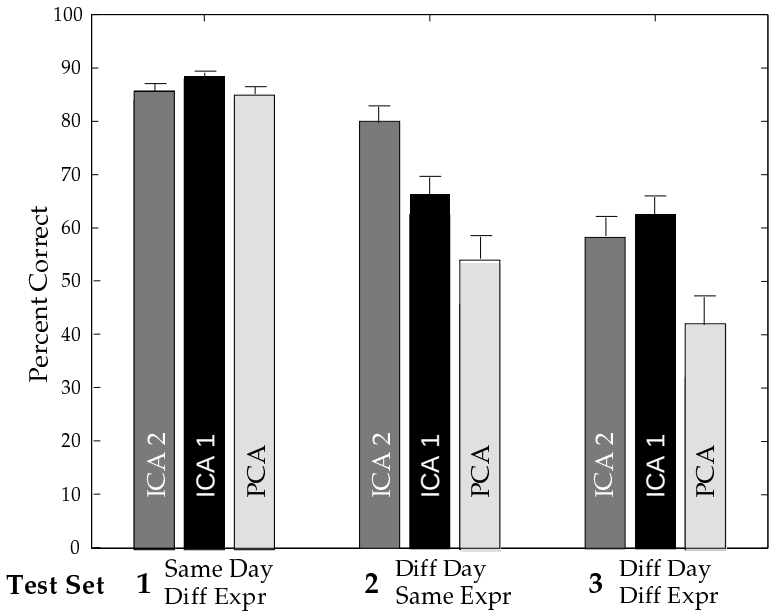

Figure 3.13. Recognition performance of the factorial code ICA representation (ICA2) using all 200 coefficients, compared to the ICA independent basis representation (ICA1), and the PCA representation, also with 200 coefficients.

difference in performance between ICA1 and ICA2 for Test Set 3 following the discriminability analysis just misses significance ($Z = 1.88, p = 0.06$).

In this implementation, we separated 200 components using 425 samples, which was a bare minimum. Test images were not used to learn the independent components, and thus our recognition results were not due to overlearning. Nevertheless, in order to determine whether the findings were an artifact due to small sample size, recognition performances were also tested after separating 85 rather than 200 components, and hence estimating fewer weight parameters. The same overall pattern of results was obtained when 85 components were separated. Both ICA representations significantly outperformed the PCA representation on Test Sets 2 and 3. With 85 independent components, ICA1 obtained 87%, 62%, 58% correct performance, respectively on Test Sets 1, 2, and 3, ICA2 obtained 85%, 76%, and 56% correct performance, whereas PCA obtained 85%, 56% and 44% correct, respectively. Again, as found for 200 separated components, selection of subsets of components by class discriminability improved the performance of ICA1 to 86%, 78%, and 65%, respectively, and had little effect on the performances with the PCA and ICA2 representations. This suggests that the results were not simply an artifact due to small sample size.

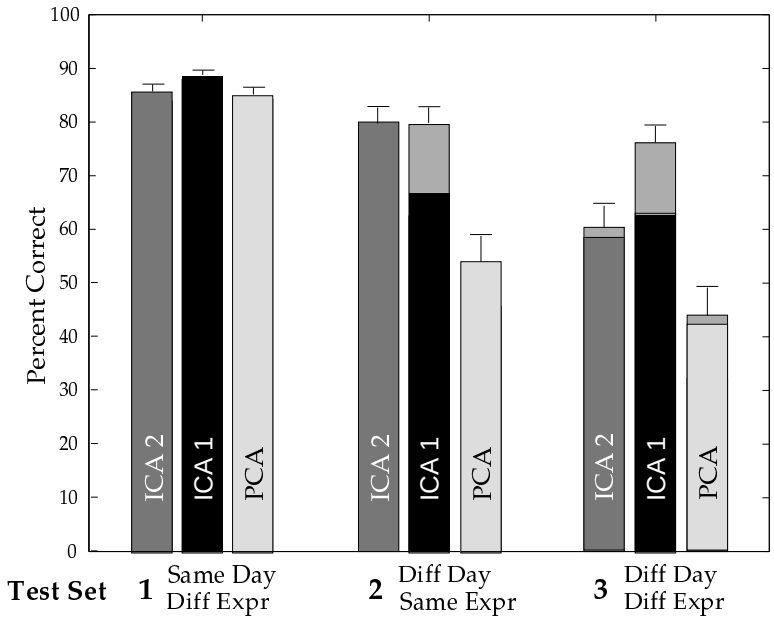

Figure 3.14. Improvement in recognition performance of the two ICA representations and the PCA representation by selecting subsets of components by class discriminability. Gray extensions show improvement over recognition performance using all 200 coefficients.

4. EXAMINATION OF THE ICA REPRESENTATIONS

4.1. Mutual information

A measure of the statistical dependencies of the face representations was obtained by calculating the mean mutual information between pairs of 50 basis images. Mutual information was calculated as

$$I(u_1, u_2) = \frac{H(u_1) + H(u_2) - H(u_1, u_2)}{H(u_1)} \tag{3.9}$$

$$\text{where } H(u_1) = -E\left[log(P_{u_1})\right].$$

Figure 3.15 (left) compares the mutual information between *basis images* for the original graylevel images, the principal component basis images, and the ICA basis images obtained in Architecture 1. Principal component images are uncorrelated, but there are remaining high order dependencies. The information maximization algorithm decreased these residual dependencies by more than 50%. The remaining dependence may be due to a mismatch between the logistic transfer function employed in the learning rule and the cumulative density function of the independent sources, the presence of sub-Gaussian

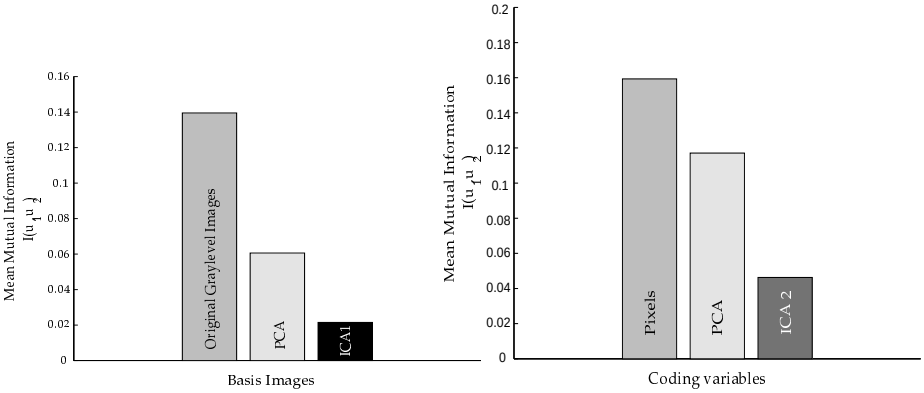

Figure 3.15. Pairwise mutual information. LEFT: Mean mutual information between basis images. Mutual information was measured between pairs of graylevel images, principal component images, and independent basis images obtained by Architecture 1. RIGHT: Mean mutual Information between coding variables. Mutual information was measured between pairs of image pixels in graylevel images, PCA coefficients, and ICA coefficients obtained by Architecture 2.

sources, or the large number of free parameters to be estimated relative to the number of training images.

Figure 3.15 (right) compares the mutual information between the *coding variables* in the ICA factorial representation obtained with Architecture 2, the PCA representation, and graylevel images. For graylevel images, mutual information was calculated between pairs of pixel locations. For the PCA representation, mutual information was calculated between pairs of principal component coefficients, and for the ICA factorial representation, mutual information was calculated between pairs of coefficients, b. Again, there were considerable high-order dependencies remaining in the PCA representation that were reduced by more than 50% by the information maximization algorithm. The ICA representations obtained in these simulations are most accurately described not as "independent," but as "redundancy reduced," where the redundancy is less than half that in the principal component representation.

4.2. Sparseness

Field (Field, 1994) has argued that sparse distributed representations are advantageous for coding visual stimuli. 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. In such a code, the redundancy of the input is transformed into the redundancy of the response patterns of the the individual outputs. Maximizing sparseness while maintaining information is equivalent to the minimum

entropy codes discussed by Barlow (Barlow, 1989). A transformation that minimizes the entropy of the individual outputs encourages statistical independence between the outputs.[3]

Given the relationship between sparse codes and minimum entropy, the advantages for sparse codes as outlined by Field (Field, 1994) mirror the arguments for independence presented by Barlow (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 units 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 how much they are active. 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).

The probability densities for the values of the coefficients of the two ICA representations and the PCA representation are shown in Figure 3.16. The sparseness of the face representations were examined by measuring the kurtosis of the distributions. Kurtosis is defined as the ratio of the fourth moment of the distribution to the square of the second moment, normalized to zero for the Gaussian distribution by subtracting 3:

$$kurtosis = \frac{\sum_i (b_i - \overline{b})^4}{\left(\sum_i (b_i - \overline{b})^2\right)^2} - 3. \tag{3.10}$$

The kurtosis of the PCA representation was measured for the principal component coefficients. The principal components of the face images had a kurtosis of 0.28. The coefficients, b, of the independent basis representation from Architecture 1 had a kurtosis of 1.25. In contrast, the coefficients, b, of the ICA factorial code representation from Architecture 2 was highly kurtotic, at 102.9.

[3]Information maximization is consistent with minimum entropy coding. By maximizing the *joint* entropy of the output, the entropies of the *individual* outputs tend to be minimized.

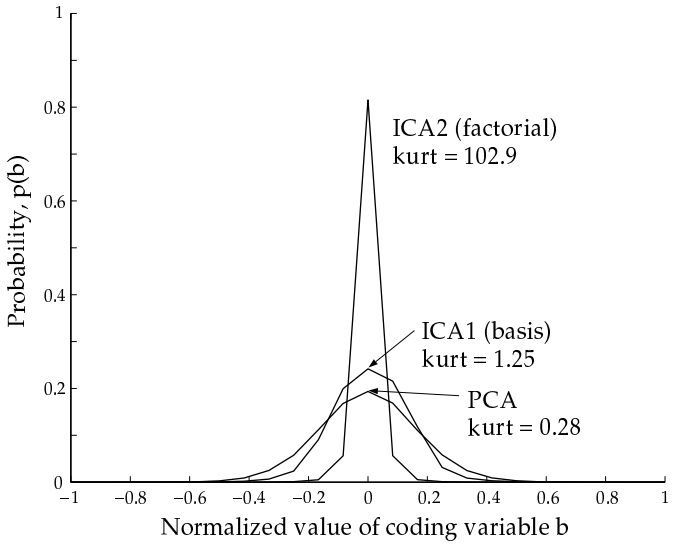

Figure 3.16. Kurtosis (sparseness) of ICA and PCA representations.

5. COMBINED ICA RECOGNITION SYSTEM

Given that the two ICA representations gave similar recognition performances, we examined whether the two representations gave similar patterns of errors on the face images. There was a significant tendency for the two algorithms to misclassify the same images. The probability that the ICA-factorial representation (ICA2) made an error given that the ICA-basis representation (ICA1) made an error was .72, .88, and .89 respectively for the three test sets. These conditional error rates were significantly higher than the marginal error rates $(Z = 7.4, p < .001; Z = 3.4, p < .001; Z = 2.8, p < .01)$, respectively. Examples of successes and failures of the two algorithms are shown in Figure 3.17.

When the two algorithms made errors, however, they did not assign the same incorrect identity. Out of a total of 62 common errors between the two systems, only once did both algorithms assign the same incorrect identity. The two representations were therefore used in conjunction to provide a reliability measure, where classifications were accepted only if both algorithms gave the same answer. This combined ICA recognition system gave an overall classification performance of 99.8% for the 400 images that met this simple criterion, out of the total of 509 test images (100%, 100%, and 97% for the three test sets, respectively).

Because the confusions made by the two algorithms differed, a combined classifier was employed in which the similarity between a test image and a

Figure 3.17. Recognition successes and failures. Left: Two face image pairs which both ICA algorithms correctly recognized. Right: Two face image pairs that were misidentified by both ICA algorithms. Images from the FERET face database were reprinted with permission from Jonathon Phillips.

gallery image was defined as $d_1 + d_2$, where d_1 and d_2 correspond to the similarity measure d in Equation 3.7 for ICA1 and ICA2 respectively. Class discriminability analysis was carried out on ICA1 and ICA2 before calculating d_1 and d_2. Performance of the combined classifier is shown in Figure 3.18. The combined classifier improved performance to 91.0% ,88.9%, and 81.0% for the three test cases, respectively. The difference in performance between the combined ICA classifier and PCA was significant for all three test sets $(Z = 2.7, p < 0.01; Z = 3.7, p < .001; Z = 3.7; p < .001)$.

6. DISCUSSION

In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels. Face representations such as "eigenfaces" and "holons" are based on principal component analysis, which separates the second-order statistics of the image set, but does not address the high-order relationships in the images. We derived two representations for face recognition based on the statistically independent components of face images. One representation used independent component analysis to find a set of independent basis images that can be considered a set of independent facial feature images. This representation was obtained by employing an architecture that found a set of independent images across spatial location. The representation defined faces as a linear combination of a set of independent feature images. The face code consisted of the coefficients for the linear combination of basis images that comprised each face image. The second representation used ICA to find a factorial face code, in which the coding variables were independent. This representation was obtained by employing an architecture that separated a set of independent coding variables across images. Both ICA representations embodied a prior that these image

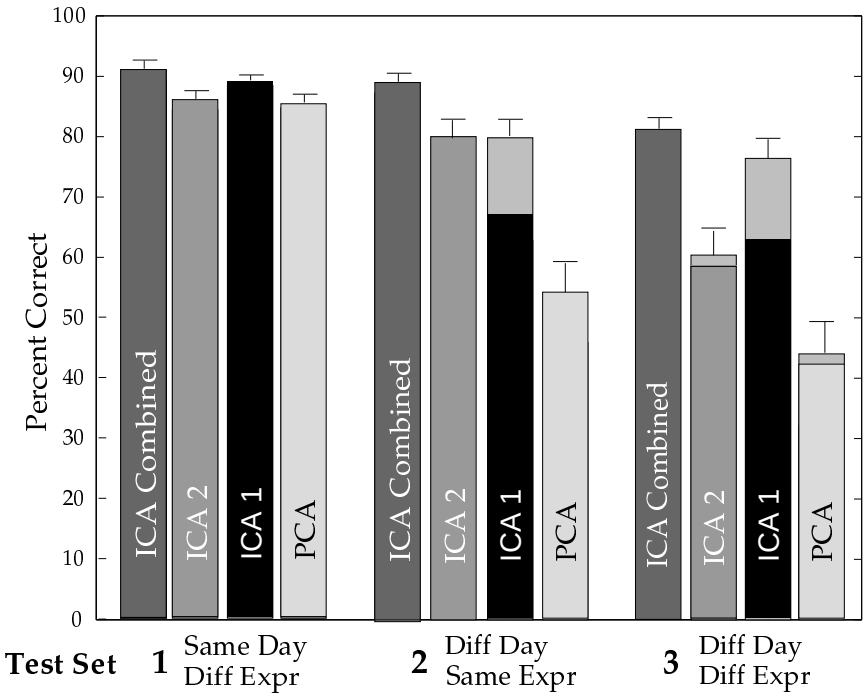

Figure 3.18. Face recognition performance of the combined ICA classifier, compared to the individual classifiers for ICA1 and ICA2, and PCA.

features were independent across individuals, so that when there were statistical dependencies, they more reliably signaled a feature combination that occurred within an individual.

Principal component analysis defines face space in terms of directions of covariance in the data. ICA, on the other hand defines face space in terms of directions of maximum dependence. ICA encodes the statistical dependencies that are expected in the input and removes them from the output. Each output unit learns a set of weights that encodes a portion of the statistical dependencies in the input, so that the dependencies are removed from between the output units.

Both ICA representations outperformed the "eigenface" representation (Turk and Pentland, 1991), which was based on principal components, for recognizing images of faces sampled on a different day from the training images. This result is particularly encouraging, since most applications of automated face recognition require identification of images collected on a different day from the sample images. These images can differ in the precise lighting conditions and facial pose, in addition to possible gross differences due to changes in hair, make-up, and facial expression. A classifier that combined the two ICA rep-

resentations outperformed eigenfaces across all three test sets. Methods have been presented for optimizing recognition performance with "eigenfaces," such as building a modular representation consisting of "eigenfaces" plus "eigenfeatures," which are principal components of subimages containing the eyes, nose, or mouth (Pentland et al., 1994). These optimization procedures are applicable to the ICA representations as well. The purpose of the comparison in this paper was to compare the ICA and PCA-based representations under identical conditions.

In Section 3, independent component analysis provided a set of statistically independent coefficients for coding the images. It has been argued that such a factorial code is advantageous for encoding complex objects that are characterized by high-order combinations of features, since the prior probability of any combination of features can be obtained from their individual probabilities (Barlow, 1989; Atick, 1992). According to the arguments of both Field (Field, 1994) and Barlow (Barlow, 1989), the ICA-factorial representation is a more optimal object representation than the ICA-basis representation given its sparse, factorial properties. Due to the difference in architecture, the ICA-factorial representation always had fewer training samples to estimate the same number of free parameters as the ICA-basis representation. Figure 3.15 shows that the residual dependencies in the ICA factorial representation were higher than in the ICA basis representation. The ICA-factorial representation may prove to have a greater advantage given a much larger training set of images. It also is possible that the factorial code representation may prove advantageous with a more powerful recognition engine than nearest neighbor on cosines, such as a Bayesian classifier. An image set containing many more frontal view images of each subject will be needed to test that hypothesis. Statistical analyses such as those conducted here require a large number of images. Performance improved as the number of components increased, where number of training samples required for learning increases multiplicatively with the number of components separated.

Unlike principal component analysis, independent component analysis using Architecture 1 found a spatially local face representation. Local feature analysis (LFA) (Penev and Atick, 1996) also finds local basis images for faces, but using second-order statistics. The LFA basis images are found by performing zero phase whitening (Equation 3.3) on the principal component axes, followed by a rotation to topographic correspondence with pixel location. The LFA algorithm is not sensitive to the high-order dependencies in the face image ensemble, and in tests to date, recognition performance with the algorithm has not significantly improved upon PCA (Donato et al., 1999).

Architecture 1 produced local basis images, but the face codes were not sparse. Architecture 2 produced sparse face codes, but with holistic basis images. A representation that has recently appeared in the literature, non-

negative matrix factorization (NMF) (Lee and Seung, 1999) produced local basis images and sparse face codes. While this representation is interesting from a theoretical perspective, it has not yet proven useful for recognition. A face representation that employs restricted Boltzmann machines (RBMs) also finds local features when non-negative weight constraints are employed (Teh and Hinton, 2001). In this novel approach, a nonlinear generative model is created for each individual and trained on pairs of different images of the that individual. Test images are compared to training images by measuring how well the generative model can account for the pair (train,test). In experiments to date, RBM's outperformed PCA for recognizing faces across changes in expression or addition/removal of glasses, but performed more poorly for recognizing faces across different days. As will be discussed in Chapter 6 it appears that spatial locality and sparseness alone are not enough for good recognition performance. Encoding high order dependencies may be the crucial property.[4]

The information maximization learning algorithm was developed from the principle of optimal information transfer in sigmoidal neurons. It contains a Hebbian correlational term between the nonlinearly transformed outputs and weighted feedback from the linear outputs (Bell and Sejnowski, 1997). The biological plausibility of the learning algorithm, however, is limited by fact that the learning rule is nonlocal. Local learning rules for independent component analysis are presently under development (Lin et al., 1997).

The principle of independence, if not the specific learning algorithm employed here (Bell and Sejnowski, 1997), may have relevance to face and object representations in the brain. Horace Barlow (Barlow, 1989) and Joseph Atick (Atick, 1992) have argued for redundancy reduction as a general coding strategy in the brain. This notion is supported by the findings of Bell and Sejnowski (Bell and Sejnowski, 1997) that the filters that produce independent outputs from natural scenes are local, oriented, spatially opponent filters similar to the response properties of V1 simple cells. Olshausen and Field (Olshausen and Field, 1996b; Olshausen and Field, 1996a) obtained a similar result with a sparseness objective, where there is a close information theoretic relationship between sparseness and independence (Barlow, 1989; Bell and Sejnowski, 1997). Conversely, it has also been shown that Gabor filters, which closely model the responses of V1 simple cells, are sensitive to high order dependencies. Gabor filter outputs to natural scenes are at least pairwise independent in the presence of divisive normalization such as the contrast gain control mechanisms proposed to exist in V1 (Simoncelli, 1997; Heeger, 1992). In further support of the biological relevance of independence, it has recently been reported that the ICA representation presented in Section 2 gave better

[4] Although the NMF codes were sparse, they were not a minimum entropy code as the objective function did not maximize sparseness while preserving information.

correspondence with human perception of facial similarity than both PCA and non-negative matrix factorization (Hancock, 2000).

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. Such mechanisms predict neural codes in both vision(Bell and Sejnowski, 1997; Wachtler et al., 2001) and audition (Lewicki and Olshausen, 1999). The research in this chapter found that face representations derived from filtes sensitive to high order dependencies gave superior recognition performance to representations derived from filters sensitive to second-order redundancies only. This finding supports arguments that independence is a good strategy for high-level object recognition.

Acknowledgments

I am grateful to Javier Movellan, Martin McKeown, and Michael Gray for helpful discussions on this topic, and valuable comments earlier drafts of this paper. Support for this work was provided by Lawrence Livermore National Laboratories ISCR agreement B291528, the McDonnell-Pew Center for Cognitive Neuroscience at San Diego, and the Howard Hughes Medical Institute.

Portions of the research in this chapter use the FERET database of facial images, collected under the FERET program of the Army Research Laboratory. An abbreviated version of this chapter appears in *Proceedings of the SPIE Symposium on Electonic Imaging: Science and Technology; Human Vision and Electronic Imaging III, Vol. 3299*, B. Rogowitz & T. Pappas, (Eds.), SPIE, 1998.

Chapter 4

AUTOMATED FACIAL EXPRESSION ANALYSIS

Abstract

The ability to recognize facial signals is essential for natural communication between humans and until recently has been absent from the computer. Within the past decade, significant advances have enabled computer systems to understand and use this natural form of human communication. Because most investigators have limited their analysis to a small set of posed expressions, the generalizability of these systems to real world applications is low. Here we present an approach to automatic facial expression analysis based on the Facial Action Coding System (FACS). This system objectively measures facial expressions by decomposing them into component actions. FACS is presently performed by expert human observers, not computers. An automated facial action coding system will have a wide range of applications in behavioral science, medicine, and human-computer interaction. This chapter reviews the state of the art in automated facial expression analysis, describes the Facial Action Coding System, and outlines our approach to automating FACS.

Facial expression is one of the most powerful and immediate means for human beings to communicate their emotions, intentions, and opinions to each other. Facial expression measurement from video is presently used in a variety of areas of behavioral research, including the study of emotion, social interaction, communication, anthropology, personality, and child development (for reviews see (Ekman et al., 1992; Ekman and Oster, 1979; Ekman and Rosenberg, 1997)). Recent advances in computer vision and neural networks open up the possibility of automatic measurement of facial signals. An automated system would have a tremendous impact on basic research by making facial expression measurement more accessible as a behavioral measure. This research also lays the basis for computer systems that can understand this important aspect of human communication. Computer systems with this capability would have a wide range of applications in education, behavioral science, mental

health, human-computer interaction, and any context in which it is important to monitor the emotional well being and paralinguistic communication of people. (See (Picard, 1997) for an in depth discussion.)

In the past decade, important progress has been made toward enabling computer systems to recognize facial expressions. The approaches that have been explored include analysis of facial motion (Mase, 1991; Yacoob and Davis, 1994; Rosenblum et al., 1996; Essa and Pentland, 1997), measurements of the shapes of facial features and their spatial arrangements (Lanitis et al., 1997; Zhang et al., 1998), holistic spatial pattern analysis using techniques based on principal component analysis (Cottrell and Metcalfe, 1991; Padgett and Cottrell, 1997; Lanitis et al., 1997), graylevel pattern analysis using local spatial filters (Padgett and Cottrell, 1997; Zhang et al., 1998), and methods for relating face images to physical models of the facial skin and musculature (Mase, 1991; Terzopoulus and Waters, 1993; Li et al., 1993; Essa and Pentland, 1997). These systems demonstrate approaches to face image analysis that are applicable to the present goals and are reviewed in Section 1, but generalizability of these systems to real world applications is low. The reasons are discussed in Section 2. (See also (Hager and Ekman, 1995).)

1. REVIEW OF OTHER SYSTEMS

1.1. Motion-based approaches

The majority of the computer vision work on facial expression recognition has focused on facial motion analysis through optic flow estimation. If the tissues and muscles are similar between different people, the motions that result from facial action should be similar, independent of surface level differences between faces. In an early exploration of facial expression recognition, Mase (1991) used optic flow to estimate the activity in 12 of the 44 facial muscles. For each muscle he defined an axis of expansion and contraction and a window in the face image within which to measure optic flow. A coarse estimate of the activity of the muscle was provided by the mean cosine of the angle of each of the flow vectors to the axis of contraction.

Yacoob & Davis (1994) constructed a mid-level representation of facial motion from the optic flow output, which consisted of such descriptions as “right mouth corner raises”. The mid-level representation was then classified into one of six facial expressions using a set of heuristic rules. Rosenblum, Yacoob & Davis (1996) expanded this work to analyze facial expressions using the full temporal profile of the expression, from initiation, to apex, and relaxation. They trained radial basis function neural networks to estimate the stage of an expression from a facial motion description, and constructed separate networks for each expression. Radial basis functions approximate

nonlinear mappings by Gaussian interpolation of examples, and are well suited to modeling systems with smooth transitions between states.

Beymer, Shashua, and Poggio (Beymer et al., 1993) trained radial basis function neural networks to learn the transformation from optic flow fields to pose and expression coordinates, and from pose and expression coordinates back to optic flow fields. The estimated optic flow fields could be used to synthesize new poses or expressions from an example image by image warping techniques.

Lien and colleagues (Lien et al., 2000) achieved some success for recognizing facial actions from motion flow fields. The flow fields were reduced to lower dimensions through principal component analysis, and then classified in a hidden Markov model.

These approaches to facial expression recognition focused exclusively on analysis of facial motion, often citing a behavioral study by Bassili (Bassili, 1979) to support the exclusive use of motion. Motion is an important aspect of facial expressions, but not the only cue. Bassili's study used point-light displays to demonstrate that human subjects *can* recognize facial expressions from motion signals alone. Recognition rates, however, were just above chance, and substantially lower than those reported for recognizing a similar set of expressions from static graylevel images (e.g. (McKelvie, 1995)). Here, we compare motion-based methods for facial expression analysis to methods that extract other forms of information from the image graylevels. We also explore combining motion with spatial texture information. Perhaps combining motion and graylevel information will ultimately provide the best facial expression recognition performance, as it does for the human visual system (Bassili, 1979; Wallbott, 1992).

1.2. Feature-based approaches

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). Measuring the positions of specific facial features can also be applied to expression recognition. Lanitis, Taylor, & Cootes, (1997), recognized identity, gender, and facial expressions by measuring shapes and spatial relationships of a set of facial features using a flexible face model. An advantage of the feature-based approach is that it drastically reduces the number of input dimensions, facilitating the classification step. 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 in

human face processing (Bruce, 1988; Bruce, 1998), and may therefore contain important information for face image analysis by computer as well.

The work most closely related to our approach to facial expression analysis is by a group lead by Jeff Cohn and Takeo Kanade. This group is developing methods independently for automating the Facial Action Coding system. (See Section 3.) Their system classifies facial actions by analysis of the locations of specific facial features and their displacements. An early version of this system (Cohn et al., 1999) explored feature point tracking, in which over 40 points were manually located in the initial face image, and the displacements of these feature points were estimated by optic flow. Discriminant functions classified the displacements into 3 action classes in the brow region, 3 in the eye region, and 9 in the mouth region. Tian, Kanade, and Cohn (Tian et al., 2001) extended this work by building multi-state facial component models to track and model facial features. These feature-based parameters were then classified in a 3 layer neural network. Here we explore image representations that provide information about entire regions of the face image, not just locations of selected feature points. In direct comparisons, techniques based on template matching, and more generally, techniques in which the image is decomposed using graylevel image kernels such as Gabor wavelet decomposition, have shown to be more effective than feature-based representations for identity recognition (Brunelli and Poggio, 1993; Lanitis et al., 1997) and expression recognition (Zhang et al., 1998).

1.3. Model-based techniques

Several facial expression recognition systems have employed explicit physical models of the face (Mase, 1991; Terzopoulus and Waters, 1993; Li et al., 1993; Essa and Pentland, 1997). Essa & Pentland (1997) extended an anatomical and physical model of the face developed by Terzopoulus and Waters (1993) and applied it to both recognizing and synthesizing facial expressions. The model consisted of a geometric mesh with 44 facial muscles, their points of attachment to the skin, and the elastic properties of the skin. Images of faces were mapped onto the physical model by image warping based on the locations of six points on the face. Motion estimates from optic flow were refined by the physical model using a Kalman filter in a recursive estimation-and-control framework, and the estimated forces were used to classify the facial expressions.

In a model-based system, classification accuracy is limited by the validity of the model. There are numerous factors that influence the motion of the skin following muscle contraction, and it is difficult to accurately account for all of them in a deterministic model. Here, we take an image-based approach in which facial action classes are learned directly from example image sequences of the actions, bypassing the physical model. Image-based approaches have recently

been advocated (Beymer and Poggio, 1996) and can successfully accomplish tasks previously assumed to require mapping onto a physical model, such as expression synthesis, face recognition across changes in pose, and synthesis across pose (Beymer et al., 1993; Vetter and Poggio, 1997).

1.4. Holistic analysis

An alternative to feature-based image analysis, holistic analysis, emphasizes preserving the original images as much as possible and allowing the classifier to discover the relevant features in the images (Movellan, 1995). An example of this approach is template matching. Templates capture information about configuration and shape that can be difficult to parameterize. In direct comparisons, template matching outperformed feature-based methods for face recognition (Brunelli and Poggio, 1993; Lanitis et al., 1997) and expression recognition (Zhang et al., 1998). The work presented in the next two chapters takes an adaptive approach to image analysis in which image features relevant to facial actions are learned directly from example image sequences, bypassing the physical model. 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).

One holistic spatial representation, "eigenfaces," is based on the principal components of the image pixels (Cottrell and Fleming, 1990; Turk and Pentland, 1991). Principal component analysis (PCA) finds an orthogonal set of dimensions that account for the principal directions of variability in the dataset. The component axes are template images that can resemble ghost-like faces. 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. Principal component analysis has been applied successfully to recognizing both facial identity (Cottrell and Fleming, 1990; Turk and Pentland, 1991), and facial expressions (Cottrell and Metcalfe, 1991; Bartlett et al., 1996; Padgett and Cottrell, 1997). Another holistic spatial representation is obtained by a class-specific linear projection of the image pixels (Belhumeur et al., 1997). Accurate alignment of the faces is critical to the success of such image-based approaches. This is true of motion-based approaches as well, however, and feature-based approaches require precise alignment of multiple internal features. Feature-based and template-based methods need not be mutually exclusive. Lanitis, Taylor, & Cootes, (1997), recognized identity, gender, and facial expressions by measuring shapes and spatial relationships of a set of facial features using a flexible face model. Performance improved by augmenting a set of feature measurements with parameters containing information

about modes of variation in graylevel images based on principal component analysis.

2. WHAT IS NEEDED

Most of the computer systems for recognizing facial expressions described above attempt to classify expressions into a few broad categories of emotion, such as happy, sad, or surprised. As support for this approach,the authors cite evidence for seven universal facial expressions (see Ekman, 1989, for a review). The existence of universal facial expressions does not imply that these seven emotion categories are sufficient to describe all facial behavior (Hager and Ekman, 1995). In natural interaction, prototypic expressions of basic emotions occur relatively infrequently. Annoyance, for example, may be indicated by just a tightening of the mouth. For real world applications, what is needed is a facial expression analysis system that is *objective, comprehensive,* and reliably linked to *ground truth.*

Comprehensiveness. If automated facial measurement were to be constructed simply in terms of seven elementary emotional categories, much important information would be lost: variations within an emotional category (eg. vengeance vs. resentment), variations in intensity (annoyance vs. fury), blends of two or more emotions (e.g. happiness + disgust $\rightarrow$ smug), facial signals of deceit, signs of cognitive state such as boredom, interest, confusion, and stress, and conversational signals that provide emphasis to speech and information about syntax.

Objectivity. For basic research into facial behavior itself, the measure of facial expression needs to be objective as well as comprehensive. Such research questions include "what are the facial signals of stress?" and "what are the differences between spontaneous and posed smiles?" A system that classifies faces as "stressed" or "not stressed" does not specify the differences in facial movement. An objective and detailed parameterization of facial movement is required for such studies. (Ekman et al., 1988). Several computer vision systems explicitly parameterize facial movement (Yacoob and Davis, 1994), and relate facial movements to the underlying facial musculature (Mase, 1991; Essa and Pentland, 1997), but it is not known whether these descriptions are sufficient for describing the full range of facial behavior. Furthermore, many of these movement parameters were estimated from posed, prototypical expressions and may not be appropriate descriptors for spontaneous facial expressions, which differ from posed expressions in both their morphology and their dynamics (Hager and Ekman, 1995) (See Section 4).

Ground truth. Finally, the system needs to be reliably linked to ground truth. If we want a system to identify "stress" in the face, how do we teach the system what a stressed face looks like? The problem seems simple at first: Ask subjects to pose an expression of stress and use a set of such images to train the computer.

A difficulty with that approach is that there are many differences between spontaneous and posed expressions, which are discussed below in Section 4. It is preferable to record subjects when they spontaneously express an emotional or cognitive state. Challenges with this approach include inducing these states and verifying that the subject is experiencing the desired state. (I have seen subjects show contempt during a "sad" film clip, and annoyance during a comedy clip.) An alternative approach is to take advantage of what is already known about how faces move during different emotions and cognitive states. There is already a large body of behavioral data in the psychology literature on facial expressions and their associations with emotional and cognitive state. One way to tackle the ground truth problem is to make use of this body of data.

3. THE FACIAL ACTION CODING SYSTEM (FACS)

We chose to base our system on the Facial Action Coding System (FACS) (Ekman and Friesen, 1978), a system employed by experimental psychologists for over 20 years to study facial behavior. FACS is a scoring system defined for expert human observers, not a computer. The system is objective, comprehensive, and there is over 20 years of behavioral data on the relation of its movement parameters to emotion and cognitive state. FACS was developed to provide objective measures of facial activity to enable behavioral science investigations of the face. Such studies included the differences in facial behavior when people are telling the truth versus lying, the patterns of central nervous system activity that accompany different facial movements, and whether facial behavior predicts clinical improvement. The difference between facial measurement as an approach to the study of facial expression versus measuring information that observers infer from facial expressions is reviewed in (Ekman, 1982a; Ekman, 1982b).

FACS was developed following extensive study of facial movement by behavioral scientists. Ekman and Friesen determined from palpation, knowledge of anatomy, and videotapes how the contraction of each of the facial muscles changed the appearance of the face (see Fig 4.1). They defined 46 Action Units, or AUs, to correspond to each independent motion of the face. (See Table 4.1.)An additional 20 actions code head and eye movements. A trained human FACS coder decomposes an observed expression into the specific AUs that produced the expression. FACS is coded from video, and the code provides precise specification of the dynamics (duration, onset and offset time) of facial movement in addition to the morphology (the specific facial actions which occur). Electromyography also directly measures facial behavior, but it is obtrusive and not comprehensive.

FACS continues to be the leading method for measuring facial expressions in behavioral science. More than 300 people worldwide have achieved intercoder agreement on the Facial Action Coding System. A number of studies

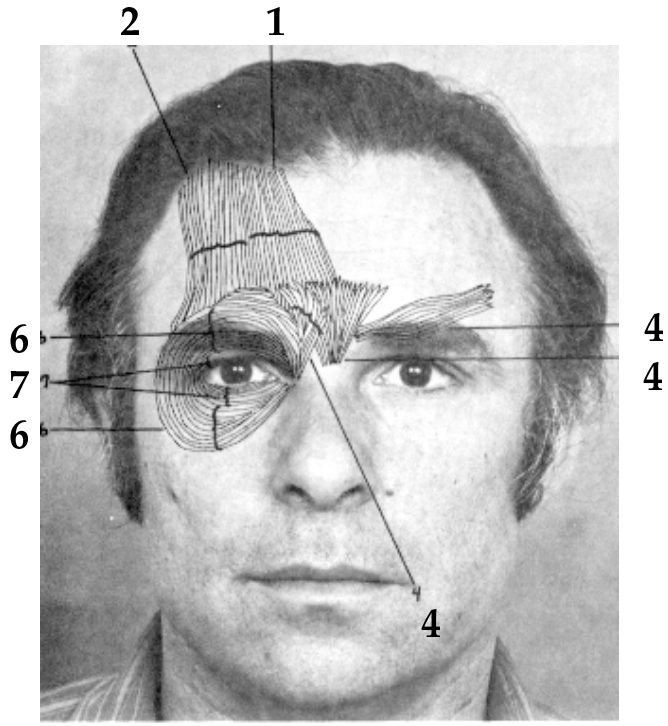

Figure 4.1. The Facial Action Coding System decomposes facial motion into component actions. The upper facial muscles corresponding to action units 1, 2, 4, 6 and 7 are illustrated. Reprinted with permission from Ekman & Friesen (1978).

have appeared showing the rich variety of information that can be obtained by using FACS (see (Ekman and Rosenberg, 1997) for a review). This system has been used, for example, to demonstrate differences between genuine and simulated pain (Craig et al., 1991), differences between when people are telling the truth versus lying (Ekman, 1985), and differences between the facial signals of suicidal and non-suicidally depressed patients (Heller and Haynal, 1994).

Aspects of FACS have been incorporated into computer graphic systems for synthesizing facial expressions (e.g. *Toy Story* (Kanfer, 1997)), and into facial muscle models for parameterizing facial movement (Rydfalk, 1987; Mase, 1991). However, FACS is not an image synthesis model. An early expression synthesis model (Rydfalk, 1987) is often called FACS in the computer vision community, which has produced some confusion. It is important to distinguish FACS itself from computer models that attempt to synthesize some of the facial actions. FACS is a scoring system for human observers. Also, although there are clearly defined relationships between FACS and the underlying facial muscles, FACS is an image-based method. Facial actions are defined by the

image changes they produce in video sequences of face images, and identified by trained human observers.

Although FACS is a promising approach, a major impediment to its widespread use in behavioral science is the time required to both train human experts and to manually score the video tape. It takes over 100 hours of training to achieve minimal competency on FACS, and each minute of video tape takes approximately one hour to score. An automated system would not only increase the speed of coding, it would also improve the reliability, precision, and temporal resolution of facial measurement. Automating the FACS would make it more widely accessible as a research tool, and it would provide a good foundation for applications of automatic facial expression analysis in man-machine interfaces. FACS provides a description of the basic elements of any facial movement, analogous to phonemes in speech. A large body of empirical data already exists demonstrating the relationship of the FACS movement parameters to emotions, emotion intensity, variations, blends, and conversational signals.

This research lays the basis for computer systems that can understand an important aspect of human communication, facial expression. Computer systems with this capability would have a wide range of applications in education, behavioral science, mental health, human-computer interaction, and any context in which it is important to monitor the emotional well being and paralinguistic communication of people. Teleconferencing and low-bit-rate encoding is another application. MPEG-4 includes facial animation parameters, which were inspired by FACS. Instead of continuously transmitting complete images of the face, one need only transmit a few movement parameters to update the image. The MPEG-4 parameters map into a subset of the action unit feature space. Some movements, such as contraction of the orbicularis oculi and wrinkling of the nose (see below), are absent from MPEG-4. FACS provides a more comprehensive parameterization of facial movement than the set of parameters presently included in MPEG-4. In addition, FACS can help with the problem of ground truth when assigning emotional or cognitive labels to a set of facial movements. For example, the MPEG-4 high-level emotion descriptor for joy is “the eyebrows are relaxed, the mouth is open, and the mouth corners are pulled back toward the ears” (MPEG Video and SNHC, 1998). This description omits the contraction of the orbicularis oculi, discussed below, and also might confuse a fear mouth, which pulls the mouth corners towards the ears using the risorious muscle, with the smiling mouth, which pulls the mouth corners towards the temples using the zygomaticus major. To date, the use of the MPEG-4 facial animation parameters has been limited to synthesis of facial expressions in the absence of an effective way to capture them from the data stream. An automated FACS system will provide this missing capability and usher in a new generation of MPEG-4 applications.

Table 4.1. List of facial actions in the Facial Action Coding System.

AU	*Name*	*Facial Muscle*
1	Inner brow raise	Frontalis, pars medialis
2	Outer brow raise	Frontalis, pars lateralis
4	Lower brows (furrow)	Corrugator supercilli, Depressor supercilli
5	Widen eye opening	Levator palpebrae superioris
6	Cheek raise	Orbicularis oculi, pars orbitalis
7	Lids tight	Orbicularis oculi, pars palpebralis
9	Nose wrinkle	Levator labii superioris alaquae nasi
10	Upper lip raise	Levator labii superioris
11	Nasolabial furrow deepen	Levator anguli oris
12	Lip corner pull	Zygomaticus major
13	Cheek puff	Zygomaticus minor
14	Dimpler	Buccinator
15	Lip corner depress	Depressor anguli oris
16	Lower lip depress	Depressor labii inferioris
17	Chin raise	Mentalis
18	Lip Pucker	Incisivii labii superioris & inferioris
19	Tongue show	Nonspecific
20	Lip stretch	Risorious with platysma
21	Neck tighten	Nonspecific
22	Lip Funnel	Orbicularis oris
23	Lip Tighten	Orbicularis oris
24	Lip Press	Orbicularis oris
25	Lips Part	Depressor labii inferioris, or relaxation of mentalis or orbicularis oris
26	Jaw drop	Masseter, relaxed temporal & internal pterygoid
27	Mouth stretch	Pterygoids and digastric
28	Lip suck	Orbicularis oris
29	Jaw thrust	Nonspecific
30	Jaw sideways	Nonspecific
31	Jaw clench	Nonspecific
32	Bite lip	Nonspecific
33	Blow	Nonspecific
34	Puff	Nonspecific
35	Cheek suck	Nonspecific
36	Tongue bulge	Nonspecific
37	Lip wipe	Nonspecific
38	Nostril Dilate	Nonspecific
39	Nostril Compress	Nonspecific
41	Lids droop	Relaxation of levator palpae superioris
42	Eyelid slit	Orbicularis oculi
43	Eyes closed	Relaxation of levator palpae superioris, orbicularis oculi, pars palebralis
44	Eye squint	Orbicularis oculi, pars palebralis
45	Eye blink	Relaxation of levator palpae superioris, orbicularis oculi, pars palebralis
46	Wink	Relaxation of levator palpae superioris, orbicularis oculi, pars palebralis

4. DETECTION OF DECEIT

Measurement of facial behavior at the level of detail of FACS can provide information for deceit detection. Investigations with FACS have revealed

a number of facial clues to deceit, including information about whether an expression is posed or genuine and leakage of emotional signals that an individual is attempting to suppress. (See (Ekman, 1991) for a compete discussion.) Spontaneous and voluntary facial expressions are mediated by different neural systems. We have very poor voluntary control over some of the facial muscles, particularly muscles in the upper face. Some facial actions tend to be omitted in posed expressions, and are less likely to be suppressed when attempting to hide an emotion (Ekman, 1991).

For example, genuine expressions of happiness can be differentiated from posed smiles by the contraction of the orbicularis oculi (AU 6) (Ekman et al., 1988). This is the sphincter muscle that circles the eye (see Figure 4.1). It raises the level of the cheek and it produces or deepens crows-feet wrinkles next to the eye. Figure 4.2a demonstrates a smile with and without the contraction of this eye muscle.

When posing an expression of sadness, subjects often exaggerate the downward turn of the mouth, and use inappropriate muscle groups. The subject on the left in Figure 4.2b demonstrates how the lip corners turn down (AU 15), without involving other muscles in the lower face. This is difficult to perform voluntarily without, for example, pushing the skin on the chin and lower lip upward (AU 17) which produces dimpling in the chin. The subject on the right was in fact doing his best to pose sadness for another researcher. This subject exaggerates the activity in the mouth region, and includes actions not associated with spontaneous expressions of sadness. He also omits activity normally observed in sadness in the upper face. The subject on the left demonstrates how the inner corners of the brow are raised (AU 1), the eyelids droop (AU 41), and the gaze is downward.

Figure 4.2c illustrates some differences between genuine and posed expressions of fear. Fear is reliably indicated by a combination of actions in the brow region in which both the inner and the outer corner of the brow is raised (AUs 1+2), and the complex of muscles between the brows is contracted (AU 4), giving the brows the raised and flat shape shown on the left in Figure 4.2c (Ekman, 1991). This combination of actions is difficult to perform voluntarily and likewise difficult to suppress when fear is experienced. The subject on the right, who is posing fear, raises the inner and outer brow as in surprise, but fails to contract the complex of muscles between the brow. This subject also omits contraction of the risorius muscle (AU 20) which pulls the lip corners towards the ear, and fails to raise the upper lid to reveal more sclera (AU 5).

Suppressed emotions can also be revealed through micro expressions. Micro expressions are full-face emotional expressions that are much shorter than their usual duration, often lasting just one-thirtieth of a second before they are suppressed or covered up with a smile (Ekman, 1991). Untrained subjects are unable to detect micro expressions when shown at full speed. An automatic

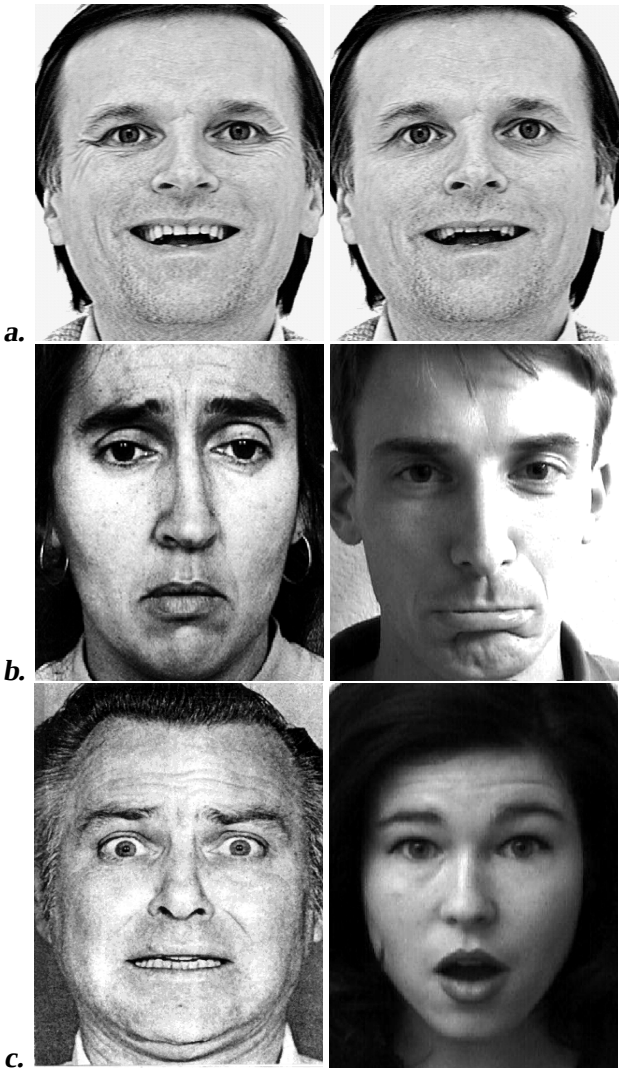

Figure 4.2. Common errors in posed facial expressions a. Genuine smiles include contraction of the sphincter muscle around the eye (left). This action is absent on the right. b. Facial behavior empirically associated with genuine sadness is demonstrated on the left. The posed expression on the right omits some actions in the upper face and includes spurious actions in the lower face. c. Spontaneous expressions of fear contain the actions shown on the left. The posed expression of fear on the right omits several actions. Courtesy of P. Ekman. Pictures of Facial Affect.

facial expression analysis system could scan large quantities of film for micro expressions in a relatively short period of time.

Other differences between spontaneous and posed expressions include symmetry. Spontaneous expressions are more symmetric than posed expressions

and have apex coordination, in which the facial muscles reach their peak contraction simultaneously (Ekman, 1991). There are also differences in the dynamics. Spontaneous expressions have a fast, smooth onset, whereas posed expressions often have a slower, jagged onset and are held too long (Ekman, 1991). There are also differences in coordination with other modalities such as timing with respect to speech.

5. OVERVIEW OF APPROACH

Chapters 5 and 6 explore methods for automating the Facial Action Coding system. Our approach to image analysis emphasizes learning and adaptive techniques. We draw upon principles from biological vision, machine learning, and probability theory to adapt processing to the immediate task environment. We separated the task of facial expression recognition into three components: (1) Face detection, tracking, and alignment; (2) Representation (often called feature extraction), and (3) Classification.

The major thrust of our work has been the investigation of component (2), image representations. What kinds of image features are useful for facial expression analysis, and which techniques are most effective for extracting information about facial expression from the image? Our work was informed by the large body of research on facial identity recognition. The pioneering work on facial identity recognition extracted relative distances between facial features such as the distance between the eyes to identify individuals (Kanade, 1973). It was later found that simple template matching techniques outperformed a detailed feature-based method for face recognition (Brunelli and Poggio, 1993). One explanation for this finding is that feature-based representations of the face can be impoverished. We do not know a priori which features and which high-order relationships among those features to measure. Our research team instead employs methods that learn about image structure directly from the image ensemble, and/or have roots in biological vision. Such methods have proven to be successful for facial identity recognition in the FERET competition (Phillips et al., 1998). Adaptive techniques in image analysis include eigenfaces (Turk and Pentland, 1991), local feature analysis (Penev and Atick, 1996), and independent component analysis (Bartlett, 1998). These methods employ image filters that are learned from the face image ensemble. These filters decompose each image into a linear superposition of basis images. Image decomposition is illustrated in Figure 4.3. Another successful technique for face recognition employs Gabor wavelet decomposition, in which the image filters are predefined, rather than learned, and are modeled after the receptive fields of primary visual cortical neurons (Lades et al., 1993; Daugman, 1988).

We applied these techniques to the problem of facial expression analysis. We compared more than ten image representation algorithms on the task of classifying facial actions. The techniques were compared on a common image

test bed using common classifiers. When comparing techniques, it is important to hold other components of the system constant, such as training and testing images, and methods for alignment and classification. Differences in the performance of feature extraction methods could be swamped, for example, by differences in alignment accuracy.

In our first comparative study (Bartlett et al., 1999), presented in Chapter 5, we explored three representations: "eigenfaces" which is an unsupervised approach to feature extraction based on pixelwise covariances (PCA), explicit feature extraction that measures facial wrinkles and eye opening, and facial motion analysis based on optic flow fields. These comparisons supported the theory that unsupervised feature extraction based on dependencies in the image ensemble is more effective for face image analysis than explicit measurement of facial features. The results also suggest that hand-engineered features plus unsupervised representations may be superior to either one alone, since their performances may be uncorrelated.

Our second study (Donato et al., 1999), presented in Chapter 6, compared adaptive filters learned through supervised and unsupervised learning to pre-defined filters based on cortical receptive fields. The representations included eigenfaces, independent component analysis, local feature analysis, Fishers linear discriminants, and Gabor wavelet decomposition. We addressed the issue of spatially local and global filters in these representations and in local implementations of principal component analysis. We also examined another motion-based representation that extracted flow fields with sub-pixel accuracy. The second study provided evidence that techniques sensitive to high-order statistical dependencies were more effective than second-order techniques for facial expression analysis.

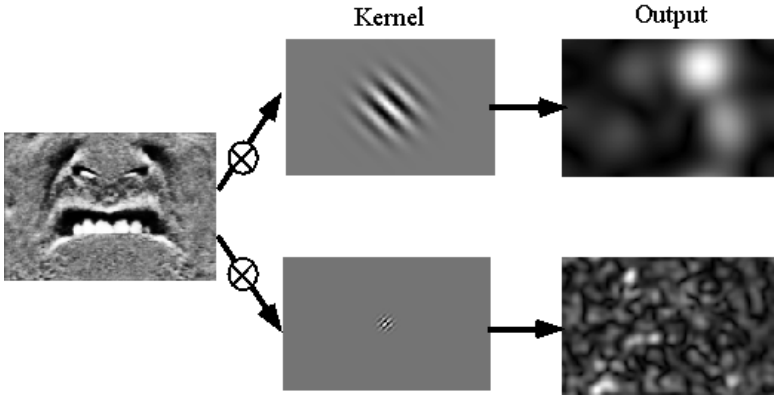

Figure 4.3. Example image decomposition. Here an image is convolved with a family of Gabor wavelets. The output is channeled to the classifier.

Chapter 5

IMAGE REPRESENTATIONS FOR FACIAL EXPRESSION ANALYSIS: COMPARATIVE STUDY I

Based on "Measuring facial expressions by computer image analysis" by Bartlett, M.S. Hager, J.C., Ekman, P., & Sejnowski, T.J., which appeared in *Psychophysiology* 36, p. 253-263, 1999. Reprinted with permission from Cambridge University Press.

Abstract

Facial expressions provide an important behavioral measure for the study of emotion, cognitive processes, and social interaction. The Facial Action Coding System, (Ekman and Friesen, 1978), is an objective method for quantifying facial movement in terms of component actions. We applied computer image analysis to the problem of automatically detecting facial actions in sequences of images. In our first study we compared three approaches: Holistic spatial analysis (eigenfaces), explicit measurement of features such as wrinkles, and estimation of motion flow fields. The three methods were combined in a hybrid system which classified six upper facial actions with 91% accuracy, including low, medium, and high magnitude facial actions. The hybrid system outperformed human non-experts on this task, and performed as well as highly trained experts. These comparisons supported the theory that unsupervised feature extraction based on dependencies in the image ensemble is more effective for face image analysis than explicit measurement of facial features.

In our first comparative study (Bartlett et al., 1999), we explored three different methods for classifying facial actions: holistic spatial analysis based on principal components (eigenfaces), a feature based approach that measures facial wrinkles and eye opening, and facial motion analysis based on template matching of optic flow fields. The performances of the three systems were compared and then combined into a single system that pools their strengths. Benchmarks for the performances of the automated systems were provided by naive and expert human subjects.

1. IMAGE DATABASE

The system was trained and tested using a database of directed facial actions. The full database contained over 1100 sequences containing over 150 distinct actions, or action combinations. The image database was obtained from 24 Caucasian subjects, 12 males and 12 females. Their ages ranged from 19 to 61 with a median of 30. 13 were experienced FACS coders, 8 had some FACS training, and 3 were naive. Each image sequence consisted of six frames, beginning with a neutral expression and ending with a high magnitude muscle contraction (Figure 5.1). The database therefore contained examples of the facial actions at low and medium magnitude as well as at high magnitude.[1] Trained FACS experts provided demonstrations and instructions to subjects on how to perform each action. Subjects were instructed to minimize rigid head motion. The selection of images was based on stop motion video coded by three experienced FACS coders certified with high inter-coder reliability. The criterion for acceptance of images was that the requested action and only the requested action was present.

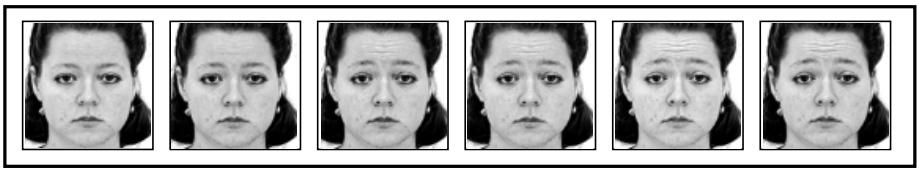

Figure 5.1. Example action sequence from the database. The example shows a subject performing AU1 starting from a neutral expression and ending with a high magnitude action.

For this investigation, we used data from 20 subjects and attempted to classify the six individual upper face actions illustrated in Figure 5.2. This set of actions was chosen for this study because the facial actions in the upper face comprise a relatively independent subset of facial actions; facial actions in the upper face have little influence on facial motion in the lower face, and vice versa (Ekman and Friesen, 1978). Most subjects were able to perform only a subset of the actions without interference from other facial muscles. Each subject performed a mean of 4 actions. The dataset therefore contained, aside from the neutral frame, a total of 400 images of facial actions (20 subjects X 4 actions X 5 frames per action). 9 subjects performed AU1, 10 performed AU2, 18 performed AU4, all 20 performed AU 5, 5 performed AU6, and 18 performed AU7.

Faces were aligned, cropped, and scaled based on the locations of two points in the first frame of each sequence. The two points were indicated by a

[1] The term "magnitude" replaces the term "intensity" used in FACS to avoid confusion with image intensity.

single mouse click at the center of each eye. All other procedures were fully automated. Accurate image registration is critical for principal components based approaches. The variance in assigned eye location using this procedure was 0.4 pixels in the 640 x 480 pixel images.

The eye positions from frame 1 were used to crop all subsequent frames, and scale the faces to 45 pixels between the eyes. The images were rotated in the plane so that the eyes were horizontal, and the luminance brightness values were linearly rescaled to [0, 255]. The images were cropped to contain only the upper half of the face, as shown in Figure 5.2. The final images contained 66 x 96 pixels. Difference images, which were used in the holistic analysis, were obtained by subtracting the neutral expression frame (the first frame in each sequence), from the five subsequent frames. Advantages of difference images include robustness to changes in illumination, removal of surface variations in facial appearance, and emphasis of the dynamic aspects of the image sequence (Movellan, 1995).

Because faces tend to be asymmetric, and the contractions of facial muscles are also frequently asymmetric, we generated additional training data by reflecting each image about the vertical axis. Mirror reversed images of *test* subjects were never included in the training set, so the classifiers had no access to information about reflected test images either during parameter estimation or classification. The reflected images were not assumed to be independent of their originals, and were not counted in the N for statistical comparisons. All 400 difference images in the dataset were asymmetric. The reflected images differed from their originals in 6125 of the 6336 pixels on average, and the mean magnitude of the difference was 5.36. Images differed *between* individuals in an average of 6179 pixels, and the mean magnitude of the difference *between* individuals was 7.17. The symmetry of the training set also ensured that the classifiers had no asymmetric bias.

2. IMAGE ANALYSIS METHODS

2.1. Holistic spatial analysis

We first evaluated the ability of a back-propagation network to classify facial actions given eigenfaces as input. This approach is based on (Cottrell and Metcalfe, 1991) and (Turk and Pentland, 1991), with the primary distinction in that we performed principal component analysis on the dataset of difference images. The remaining variation in the dataset of difference images was that due to the facial dynamics. Each of the 800 difference images was converted to a vector by concatenating the rows of pixel intensities. Hence, each image was represented as a point in a high dimensional space given by the grayvalue at each of the 6336 pixel locations. The principal component axes of the difference image data were then calculated by finding the eigenvectors of the

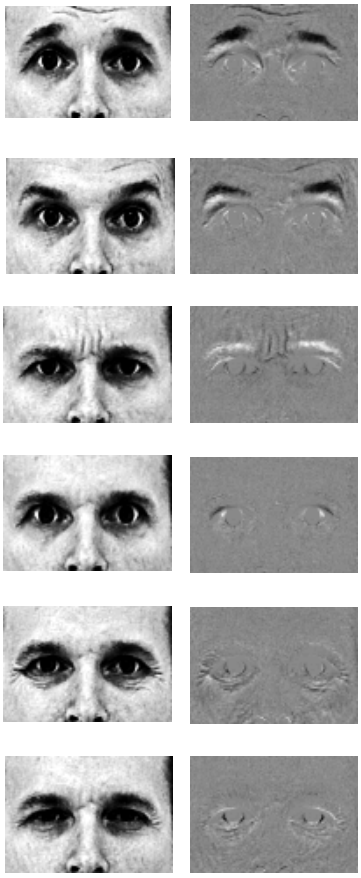

Figure 5.2. Examples of the six actions used in this study. From left to right: Cropped image of the action at highest magnitude; Difference image obtained by subtracting the neutral image (frame 1 of the sequence); Action unit number; Action unit description adapted from Ekman & Friesen (1978).

pixelwise covariance matrix. The axes were ordered by the magnitude of the corresponding eigenvalue. Figure 5.3 shows the first 12 principal components of the difference images.

The principal component representation consisted of a set of coefficients obtained by projecting each difference image onto the component axes. These coefficients comprised the input to a 2 layer neural network with 10 hidden units, and six output units, one per action. The network was feedforward, with each unit connected to all of the units in the layer above (see (Haykin, 1994)). The activities of the hidden and output units were calculated sequentially as the weighted sum of their inputs and passed through a hyperbolic tangent transfer function. The network was trained by back-propagation of error to output

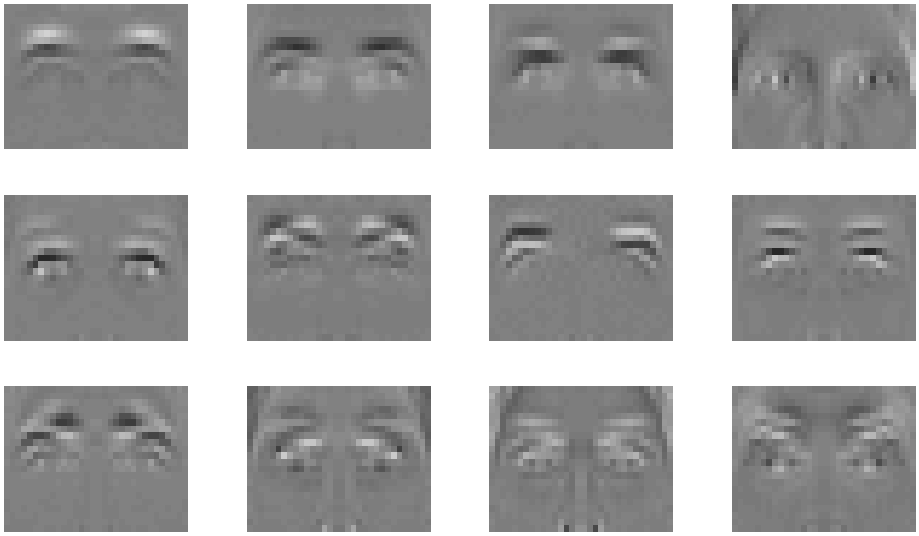

Figure 5.3. First 12 principal components of the dataset of difference images, ordered left to right, top to bottom. The first component appears to code for vertical brow position. The sixth component axis appears to differentiate between AU1, raising the inner corners of the brow, and AU2, raising the lateral portions of the brows. Component 7 appears to be an axis of left-right asymmetry in the lateral brow movement, and component 5 appears to be an eye opening axis.

a 1 for the appropriate action, and zeros everywhere else, using conjugate gradient descent on the summed squared error. Stopping criterion was the inflection point in the mean test error. The output unit with the highest activity determined the classification.

2.2. Feature measurement

Four of the upper face actions produce wrinkles in distinct locations on the face, and the remaining two alter the amount of visible sclera. We applied a method developed by Jan Larsen (Bartlett et al., 1996) for measuring changes in facial wrinkling and eye opening. The feature measurements were carried out on 360 x 240 pixel images. Facial wrinkles were measured at the four facial positions shown in Figure 5.4a, which were located in the image automatically from the eye position information. These image locations were selected for detecting wrinkles produced by AUs 1,2,4, and 6. At each location, mean pixel intensities of a five pixel wide segment were extracted and then smoothed lengthwise by a median filter. Figure 5.4b shows the smoothed pixel intensities along the image segment labeled A. The pixel intensities drop sharply at the two major wrinkles.

We chose as a measure of facial wrinkling an estimate of the sum squared derivative of the pixel intensities along the segment. This value is estimated by

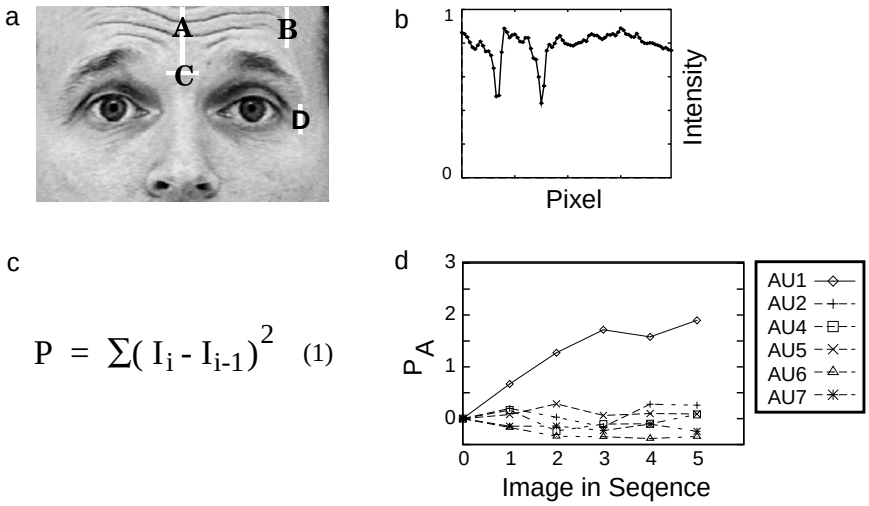

Figure 5.4. a) Wrinkling was measured at four image locations, A-D. b) Smoothed pixel intensities along the line labeled A. c) The wrinkle measure, P. I_i is the intensity of the ith pixel of the segment. d) P measured at image location A for one subject performing each of the six actions.

P (Figure 5.4c.) Pixel differences approximate the derivative (Jain et al., 1995). This measure is sensitive to both the deepening of existing wrinkles and the addition of new wrinkles. To control for permanent wrinkles, P values for the neutral image were subtracted. Figure 5.4d shows P values along line segment A, for a subject performing each of the six actions. The P values remain at zero except for AU 1, for which it increases as action magnitude increases. Only AU 1 produces wrinkles in the center of the forehead.

For detecting and discriminating AUs 5 and 7, we defined an eye opening measure as the area of visible sclera lateral to the iris. This area was found by starting at the pupil and searching laterally for connected rows of pixels above threshold. Again, differences from baseline were measured. A three-layer neural network was trained to classify each image from the five feature measures, consisting of the wrinkle feature measured at 4 locations and the eye opening measure. The network had 15 hidden units and six output units.

2.3. Optic flow

Local estimates of motion in the direction of the image gradient were obtained by an algorithm based on the brightness constraint equation (Horn and Schunk, 1981):

$$\frac{dI(x,y,t)}{dt} = \frac{\partial x}{\partial t}\frac{\partial I(x,y,t)}{\partial x} + \frac{\partial y}{\partial t}\frac{\partial I(x,y,t)}{\partial y} + \frac{\partial I(x,y,t)}{\partial t} = 0 \qquad (2)$$

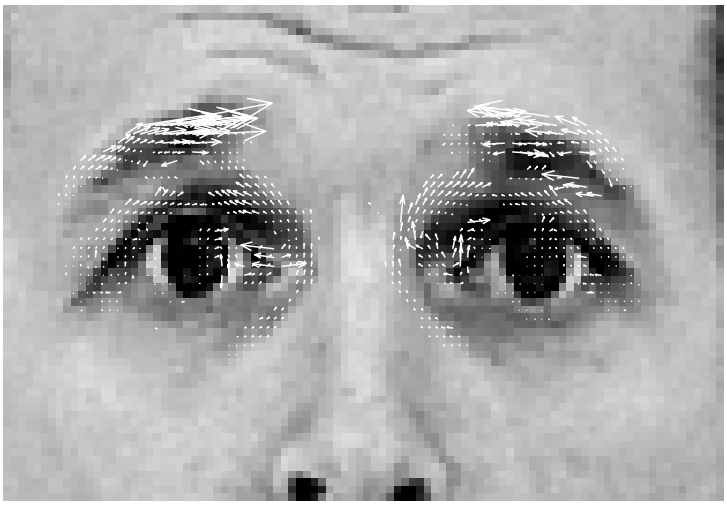

Figure 5.5. Example flow field of a subject performing AU1, inner brow raiser. The flow vector at each image location is plotted as an arrow with length proportional to the local estimate of velocity.

This equation assumes that there is no overall gain or loss of brightness in the image I over time, and any changes in brightness can be accounted for by shifts in spatial position. The local image velocities, $v_x = \frac{\partial x}{\partial t}$ and $v_y = \frac{\partial y}{\partial t}$, are defined in terms of the spatial and temporal gradients of the image, $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, and $\frac{\partial I}{\partial t}$.

Optic flow was estimated between image pairs, a given frame in an action sequence, t_i, and the neutral frame, t_0. Images were first smoothed by a 5 x 5 Gaussian kernel. Estimates of the spatial gradients, ΔI_x and ΔI_y, were obtained with horizontal and vertical Sobel edge filters. The temporal gradient was estimated by $\Delta I_t = I(x, y, t_i) - I(x, y, t_0)$. Local estimates of image velocity in the direction of the gradient were obtained by $v_x = \frac{\Delta I_t}{\Delta I_x}$ and $v_y = \frac{\Delta I_t}{\Delta I_y}$.

Gradient-based techniques for estimating optic flow give reliable estimates only at points where the gradient is high (ie. at moving edges). Velocity estimates were set to zero at locations at which the total edge measure $r = {\Delta I_x}^2 + {\Delta I_y}^2$ was beneath a threshold of 0.2. An example flow field is shown in Figure 5.5. One of the advantages of this simple local estimate of flow was speed. It took 0.13 seconds on a 120 MHz Pentium to compute one flow field.

The flows fields were classified by a template matching procedure. A weighted template for each of the actions was calculated from the training images as the mean flow field at medium action magnitude (frame 4 of the sequence). Novel flow patterns, f^n, were compared to the template f^t by the

correlational similarity measure S:

$$S(f^n, f^t) = \frac{\sum_i f_i^n \cdot f_i^t}{\sqrt{\sum_i f_i^n \cdot f_i^n}\sqrt{\sum_i f_i^t \cdot f_i^t}} \qquad (3)$$

where i indexes image location. $S(f^n, f^t)$ is the cosine of the angle between the two flow vectors.

2.4. Human subjects

2.4.1 Naive human subjects

One benchmark for the performances of the automated systems was provided by the ability of naive human subjects to classify the same images. Subjects were nine adult volunteers with no prior knowledge of facial expression measurement. Subjects were provided with a guide sheet similar to Figure 5.2 which contained an example image of each of the six actions along with a written description of each action and a list of image cues for detecting and discriminating the actions from Ekman and Friesen (1978). Each subject was given a training session in which the facial actions were described and demonstrated, and the image cues listed on the guide sheet were reviewed and indicated on the example images. The subjects kept the guide sheet as a reference during the task.

Face images were cropped and scaled identically as they had been for the automated systems, with 45 pixels between the eyes, and printed using a high resolution HP Laserjet 4si printer with 600 dpi. Because the automated systems had information about the test image and the neutral image only when making a classification, face images were presented to the human subjects in pairs, with the neutral image and the test image presented side by side. Subjects were instructed to compare the test image with the neutral image and decide which of the actions the subject had performed in the test image. Subjects were given a practice session with feedback consisting of one example of each action at high magnitude. Neither the practice face nor the reference face was used for testing. The task contained ninety-six image pairs, consisting of low, medium, and high magnitude examples of the six actions from six different faces, three male and three female. Subjects were allowed to take as much time as they needed to perform the task, which ranged from 30 minutes to one hour.

2.4.2 Expert coders

A second benchmark was provided by the agreement rates of expert coders on these images. Subjects were four certified FACS coders. The task was identical to the naive subject task with the following exceptions: Expert subjects were not given a guide sheet or additional training, and the complete face was visible, as it would normally be during FACS scoring. One hundred and fourteen image

pairs were presented, consisting of low, medium, and high action magnitude examples of the six actions from seven faces. Time to complete the task ranged from 20 minutes to one hour and 15 minutes.

3. RESULTS

Generalization to novel faces was tested using leave-one-out cross-validation (Tukey, 1958). This procedure makes maximal use of the available data for estimating parameters. System parameters were estimated 20 times, each time using images from 19 subjects for training and and reserving all of the images from one subject, including the reflected images, for testing. The system parameters were deleted and re-estimated for each test. Mean classification performance across all test images in the 20 cross-validation runs was then calculated.

Under this procedure there were 800 test images, containing low, medium and high magnitude examples of the facial actions. The systems classified the test images one frame at a time, without reference to previous outputs. Figure 5.6 plots the overall mean performances of the classifiers on novel faces. Performances by facial action are the diagonal entries in the confusion matrices in Tables 5.1 and 5.2.

Holistic spatial analysis

Classification performance was evaluated for two scales of difference images, 66 x 96 and 22 x 32, and for five quantities of principal components in the network input: 10, 25, 50, 100, and 200. There was a trade-off between increasing the amount of information in the input and increasing the number of free parameters to be estimated. The higher principal components may also include more information on between subject variations. We obtained the best performance of 88.6% using the first 50 principal components of the 22 x 32 difference images.

The holistic system with 50 principal components had 580 parameters, while our training set in a given training run contained on average 760 images. Over-parameterization is a risk with such high dimensional networks. Performance for generalization to novel faces provided a measure of how well the system performed the general class discrimination, as opposed to finding a trivial solution that minimized the error for the training samples without learning the class discrimination.

The performance of 88.6% is substantially higher than the 70% performance reported by Padgett & Cottrell (1997) for facial expression classification using full-face Eigenfaces. The success of the present system could be attributable to reduced variability due to the use of difference images, or to the smaller original image size, so that 50 principal components accounted for a greater percentage of the variability. In addition, we employed a region of interest analysis,

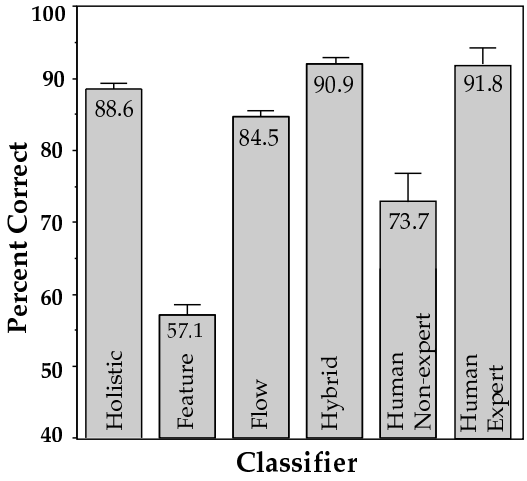

Figure 5.6. Performance comparisons for generalization to novel subjects. Values are percent correct across all test images. Error bars are one standard deviation of the estimate of the success rate in a Bernoulli distribution. Human results were prorated by action and action magnitude to match the proportions in the complete image set.

consisting of half of the face image, which is similar to the "Eigenfeature" approach that gave Padgett & Cottrell better performance.

Feature measurement

The performance of the feature-based classifier on novel faces was lower than the other methods, at 57% correct. Normalization of the feature measures with Z-scores did not improve performance. The classifier was most accurate for the two actions that involved changes in eye opening, AU5 and AU7, at 74% and 62% correct respectively. The poor performance for novel faces may be attributable to the differences in facial wrinkling patterns between subjects depending on skin elasticity, facial structure, and fat stores. The feature-based classifier performed well for new images of a face used for training, with classification accuracy of 85.3%.

Optic flow

Template matching of motion flow fields classified the facial actions with 84.5% accuracy for novel subjects. The performance of the motion-based classifier was similar to that of the holistic classifier, giving highest accuracy for AUs 2. 4, 5, and 7, and lowest for AUs 1 and 6.

3.1. Hybrid system

We obtained the best performance when we combined all three sources of information into a single neural network. The classifier was a feed forward network with 10 hidden units taking 50 component projections, 5 feature measures, and 6 template matches as input. The hybrid system improved the generalization performance to 90.9%, over the best individual method at 88.6%. While the increase is small, it constitutes about 20% of the difference between the best individual classifier and perfect performance.

We examined how the hybrid system benefited from the multiple sources of input information by looking at correlations in the performances of the three individual classifiers. The contribution of additional inputs to the signal-to-noise ratio depends on their correlations. Each data point in the correlation was mean percent correct for one of the twenty faces, across all actions and action magnitudes. The performances of the holistic and the flow field classifiers were correlated ($r^2 = 0.36, t(18) = 2.96, p < 0.01$). The feature-based system was not correlated with either the holistic or flow field classifiers ($r^2 = 0.05, t(18) = 0.85, p > 0.4$) and ($r^2 = 0.02, t(18) = 0.65, p > 0.5$), respectively. Although the stand-alone performance of the feature-based system was low, it contributed to the hybrid system by providing estimates that were uncorrelated with the two template-based systems. Without the feature measures, 17% of the improvement was lost.

Table 5.1. Confusion Matrix for Naive and Expert Human Subjects. Rows give the percent occurrence of each response for a given action. Nv: Naive subject data, Ex: Expert subject data.

Action	***Responses***											
	AU1		AU2		AU4		AU5		AU6		AU7	
	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex
AU1	***.84***	***.99***	.08	.00	.03	.00	.02	.00	.02	.00	.02	.01
AU2	.12	.04	***.83***	***.93***	.00	.00	.03	.00	.01	.00	.00	.02
AU4	.03	.00	.03	.01	***.88***	***.96***	.01	.00	.02	.00	.03	.02
AU5	.09	.00	.20	.01	.00	.01	***.64***	***.98***	.03	.00	.03	.01
AU6	.04	.00	.03	.01	.04	.00	.00	.00	***.55***	***.41***	.34	.58
AU7	.00	.00	.04	.00	.05	.02	.00	.00	.26	.09	***.65***	***.89***

Human subjects

A benchmark for the performance of the automated systems was provided by the performance of naive human subjects on the same set of images with identical cropping and scaling. Human non-experts classified the images with 73.7% accuracy. This is a difficult classification problem that requires considerable

training for people to be able to perform well. Performance of the naive human subjects was significantly lower than that of the hybrid system on the subset of images used in the human study ($Z = 2.04, p < 0.05$).

A second benchmark was provided by the agreement rates of expert coders on these images. The expert human subjects classified the actions with 91.8% agreement with the class labels assigned during database collection, which is well above the FACS inter-coder agreement standard for proficiency. The majority of the disagreement was on the low magnitude examples of the actions, and the absence of video motion could account for much of the disagreement. Because the images were originally labeled by two expert coders with access to stop-motion video, this data provides a measure of inter-coder *agreement* between coding stop-motion video and static images. The performance of the holistic and hybrid computer systems did not differ significantly from that of the human experts (Z = 1.63; Z = 1.86), but the expert coders did outperform the optic flow and feature-based classifiers ($Z = 3.17, p < 0.01$) and ($Z = 7.2, p < 0.001$).

3.2. Error analysis

The action confusions made by both naive and expert human subjects are presented in Table 5.1. Naive subjects made the most confusions between AUs 6 and 7, which both alter the appearance underneath the eye, followed by AUs 2 and 5, which both give an eye widening appearance by raising the outer brows and the upper lid respectively, followed by AUs 1 and 2, which raise the inner and outer portions of the eyebrows, respectively. The majority of the disagreements for the experts were between AUs 6 and 7.

Table 5.2 shows the action confusions made by the three image analysis systems and the hybrid system. Correlations among the action confusions are given in Table 5.3. Consistent with the performance rate comparisons, the confusions made by the holistic system were highly correlated with those of the motion-based system, whereas the confusions made by the feature-based system were less correlated with those of the holistic system, and uncorrelated with those of the motion-based system.

Of the four automated systems, the holistic system had the most similar pattern of confusions to both the naive human subjects and to the expert coders. This finding is consistent with previous reports that principal component representations of face images account well for human perception of distinctiveness and recognizability of faces (O'Toole et al., 1994; Hancock et al., 1996). The confusions of the feature-based system were least correlated with those of the human subjects, with a low but significant correlation with the expert coders, and no significant correlation with the naive subjects.

Table 5.2. Confusion Matrix for the automated classifiers. Rows give the percent occurrence of each response for a given action. Hol: Holistic, Mt: Motion, Ft: Feature, Hyb: Hybrid.

Action	***Responses***														
	AU1				AU2				AU4				AU5				AU6				AU7			
	Hol	Mt	Ft	Hyb	Hol	Mt	Ft	Hyb	Hol	Mt	Ft	Hyb	Hol	Mt	Ft	Hyb	Hol	Mt	Ft	Hyb	Hol	Mt	Ft	Hyb
AU1	***.58***	***.20***	***.50***	***.57***	.19	.31	.04	.17	.00	.00	.29	.01	.10	.33	.14	.08	.03	.00	.00	.00	.10	.15	.02	.18
AU2	.12	.02	.10	.10	***.83***	***.94***	***.36***	***.85***	.01	.00	.04	.00	.01	.02	.41	.00	.00	.00	.00	.00	.03	.02	.09	.05
AU4	.00	.00	.08	.00	.00	.01	.01	.00	***.96***	***.97***	***.54***	***.99***	.00	.00	.26	.00	.06	.00	.00	.00	.04	.02	.10	.01
AU5	.01	.00	.07	.00	.15	.00	.35	.00	.00	.00	.10	.00	***.98***	***1.0***	***.74***	***1.0***	.00	.00	.00	.00	.00	.00	.06	.00
AU6	.00	.00	.00	.00	.00	.00	.02	.00	.00	.16	.00	.02	.06	.04	.20	.02	***.56***	***.40***	***.38***	***.74***	.38	.40	.40	.22
AU7	.00	.00	.06	.00	.00	.00	.03	.00	.01	.00	.06	.01	.00	.02	.21	.01	.00	.03	.03	.00	***.99***	***.94***	***.62***	***.98***

Table 5.3. Action confusion correlations. Entries are squared correlation coefficients. Stars indicate statically significant correlation based on a t-test with 28 degrees of freedom at the .05* level, .01**, and .001***.

	Expert	Holistic	Motion	Feature	Hybrid
Naive	.58***	.36**	.18*	.05	.19*
Expert		.66***	.36**	.23**	.36**
Holistic			.70***	.17*	.82***
Motion				.09	.69***
Feature					.07

4. DISCUSSION

Facial action codes provide a rich description of facial behavior that enables investigation of the relationship of facial behavior to internal state. We developed methods for automatically classifying facial actions from image sequences. The approach presented here differed from other computer facial expression analysis systems in that we focused on classifying the basic elements that comprise complex facial movements rather than classifying emotion categories. Classification was learned directly from images of facial actions without mediation of a physical model.

We compared the performance of three diverse approaches to processing face images for classifying facial actions: holistic spatial analysis, feature measurement, and analysis of motion flow fields. Best performance of 91% correct for classifying 6 actions was achieved by combining the three methods of image analysis in a single system. The hybrid system classified an image in less than a second on a 120 MHz Pentium. Our initial results are promising since some of the upper facial actions included in this study require extensive training for humans to discriminate reliably. The holistic and hybrid automated systems outperformed human non-experts on this task, and the hybrid system performed as well as highly trained experts.

The image analysis methods did not depend on the precise number of video frames, nor that the actions be of any particular magnitude beyond the neutral frame. For applications in which neutral images are unavailable, principal component analysis could be performed on the original graylevel images. Methods based on principal component analysis have successfully classified static graylevel images of facial expressions (Padgett and Cottrell, 1997). The image analysis also required localization of the face in the image. For this study, the localization was carried out by making two mouse clicks, one at the center of

each eye, in the first frame of the sequence. All other aspects of the systems were fully automated. Highly accurate eye location algorithms are available (eg. Beymer, 1994), and automating this step is a realistic option. The image alignment procedure ignored out-of-plane rotations, which could be handled by methods for estimating the frontal view of a face from a non-frontal view, eg. (Beymer et al., 1993; Vetter and Poggio, 1997).

There are 46 action units, of which we have presented classification results for 6. The holistic and motion-based systems are not specific to particular actions, and can be applied to any other facial motion. The image analysis in these systems was limited to the upper half of the face because upper facial actions have little effect on motion in the lower face, and vice versa (Ekman and Friesen, 1978). We are presently applying these techniques to images of the lower half of the face to classify the lower facial actions as well.

It remains an empirical question to determine whether this approach will have the same success when dealing with spontaneous rather than deliberately made facial actions. While the morphology of the facial actions should not differ in spontaneous as compared to deliberate facial actions, the timing of the activity and the complexity of facial actions may well be different. Evaluating spontaneous facial movement is an important next step.

Most automatic facial expression analysis systems have focused on either motion or surface graylevels, but not both. It should be noted that while human subjects *can* recognize facial expressions from motion signals alone (Bassili, 1979), recognition rates are only just above chance. Likewise, although humans can recognize facial expressions quite well from static graylevel images, expression recognition improves with motion information (Wallbott, 1992). The system presented here integrates both analysis of surface graylevels and motion information.

Related work at the Univerisity of Pittsburgh and Carnegie Mellon supports the feasibility of automating the Facial Action Coding System. The approach at Pittsburgh and Carnegie Mellon has primarily emphasized feature-based representations. An early version of this system (Cohn et al., 1999) explored feature point tracking of a set of points in the face image that were initially located by hand and tracked with optic flow. Tian, Kanade, and Cohn (Tian et al., 2001) extended this work by building multi-state facial component models to track and model facial features. Here we explore adaptive image features in addition to hand-engineered ones. Although both systems rely on manual initialization, fully automating our system may be more straightforward, as it requires accurate head tracking only, and does not depend on precise localization of multiple internal features. Tian and colleagues (Tian et al., 2001) compared performance of their system to ours using the same database and the same set of six individual upper facial actions. They achieved 89.4% correct for classifying high magnitude facial actions only, which was similar to the 90.9%

reported here for classifying low and medium in addition to high magnitude facial actions. Importantly, they demonstrated that a multilayer neural network learns robust generalization to new action unit combinations.

We found that the two template-based methods, holistic spatial analysis and motion analysis, outperformed the feature-based method for facial action recognition. This supports previous findings that template approaches outperformed feature-based systems for recognizing facial identity (Brunelli and Poggio, 1993; Lanitis et al., 1997) and expression (Zhang et al., 1998). This result is also supported by Lien (2000), who found that facial furrow measurement based on analysis of high image gradients did not perform as well as full field motion analysis for facial action classification.

Our results also suggest that hand-engineered features plus templates may be superior to either one alone, since their performances may be uncorrelated. Classification of local feature measurements is heavily dependent on exactly which features were measured. Padgett & Cottrell (1997) found that local principal component analysis was superior to full-face eigenfaces for expression recognition. These local features were based on data-driven kernels obtained from the graylevels of the face images, as opposed to the hand-engineered feature measures that performed poorly in this study and others (e.g. Brunnelli & Poggio, 1993). The next chapter explores local representations of faces based on the outputs of local filters such as Gabor wavelets and local principal component analysis for facial action classification.

A completely automated method for scoring facial actions in images would make facial expression measurement more widely accessible as a research tool in behavioral science, medicine, and psychophysiology. Facial action codes have already proven a useful behavioral measure in studies of emotion (e.g. Ekman, 1984), human interaction and communication (Ekman and Oster, 1979), cognition (Zajonc, 1984), and child development (Camras, 1977). Measurement of observable facial behavior has been combined with simultaneous scalp EEG in the study of physiological patterns associated with emotional states (Davidson et al., 1990), and with measures of autonomic nervous system activity to study the relationship of emotion to facial muscles and the autonomic nervous system (Ekman et al., 1983).

Neuropsychological investigations in humans and physiological recordings in primates have indicated a separate neural substrate for recognizing facial expression independent of identity (Tranel et al., 1988; Adolphs et al., 1995; Hasselmo et al., 1989), and there is evidence that the recognition of specific facial expressions depends on distinct systems (Adolphs et al., 1996). Neural substrates for the perception of two negative emotions, fear and disgust, have recently been differentiated using fMRI (Phillips et al., 1997). Whereas perception of expressions of fear and anger produced activation in the amygdala (Breiter et al., 1996; Morris et al., 1996), perception of disgust in others acti-

vated interior insular cortex, an area involved in responses to offensive tastes (Yaxley et al., 1988; Kinomura et al., 1994).

Automated facial action coding would provide an objective measure of visual stimuli in such investigations of the neural substrates for the perception of facial expressions, as well as providing a behavioral measure of emotional state. An automated system would improve the reliability, precision, and temporal resolution of facial measurement, and would facilitate the use of facial measurement in psychophysiological investigations into the neural systems mediating emotion.

Acknowledgements

This research was supported by NSF Grant No. BS-9120868, Lawrence Livermore National Laboratories Intra-University Agreement B291436, and Howard Hughes Medical Institute. We are indebted to FACS experts Harriet Oster, Linda Camras, Wil Irwin, and Erika Rosenberg for their time and assistance. We thank Gianluca Donato, Jan Larsen, and Paul Viola for contributions to algorithm development, Wil Irwin and Beatrice Golomb for contributions to project initiation, and Claudia Hilburn Methvin for image collection. Thanks to Gary Cottrell for valuable comments on earlier drafts of this paper.

Chapter 6

IMAGE REPRESENTATIONS FOR FACIAL EXPRESSION ANALYSIS: COMPARATIVE STUDY II

Based on "Classifying Facial Actions" by G.L. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, which appeared in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(10) p. 974-989, 1999. Reprinted with permission from the IEEE.[1]

Abstract

The Facial Action Coding System (FACS) (Ekman and Friesen, 1978) is an objective method for quantifying facial movement in terms of component actions. This system is widely used in behavioral investigations of emotion, cognitive processes, and social interaction. The coding is presently performed by highly trained human experts. This chapter explores and compares techniques for automatically recognizing facial actions in sequences of images. These techniques include analysis of facial motion through estimation of optical flow; holistic spatial analysis such as principal component analysis, independent component analysis, local feature analysis, and linear discriminant analysis; and methods based on the outputs of local filters, such as Gabor wavelet representations, and local principal components. Performance of these systems is compared to naive and expert human subjects. Best performances were obtained using the Gabor wavelet representation and the independent component representation, both of which achieved 96% accuracy for classifying twelve facial actions of the upper and lower face. The results provide converging evidence for the importance of local filters, high spatial frequencies, and high-order dependencies for classifying facial actions.

1. INTRODUCTION

This chapter presents a survey and comparison of techniques for facial expression recognition as applied to automated FACS encoding. This chapter extends the comparison in Chapter 5 to include a more robust optic flow algorithm (Singh, 1991). In addition, we applied a number of methods that have appeared in the identity recognition literature to the problem of facial expression analysis. These include Gabor wavelets (Daugman, 1988; Lades et al., 1993), linear discriminant analysis (Belhumeur et al., 1997), local feature analysis (Penev and Atick, 1996), and independent component analysis (Bartlett and Sejnowski, 1997; Bartlett et al., 1998). The techniques are compared on a single image testbed. The testbed includes six lower-face actions in addition to the six upper-face actions used in Chapter 5, for a total of 12. The analysis focuses on methods for face image representation (generation of feature vectors) and the representations are compared using a common similarity measure and classifier.

Overview

Motion is an important source of information for facial expression recognition. The previous chapter explored a fast but crude optic flow technique based on image gradients. Here, in Section 3, we implement a correlation-based method with sub-pixel accuracy (Singh, 1991). Because local smoothing is commonly imposed on flow fields to clean up the signal, we also examine the effects of local smoothing on classification of facial motion.

Holistic spatial analysis is an approach that employs image-dimensional graylevel texture filters. Many of these approaches employ data-driven kernels learned from the statistics of the face image ensemble. The previous chapter explored one such representation: eigenfaces. Here, in Section 4, a number of holistic representations are examined. We compare three techniques in which the image filters are derived from unsupervised learning: Eigenfaces employs principal component analysis (PCA) which which is an unsupervised learning method based on the second-order dependencies among the image pixels (pixelwise covariances). Local feature analysis (LFA) is a related technique that is also based on principal component analysis and includes a manipulation that produces spatially local filters. Eigenfaces and LFA are insensitive to the high-order dependencies among the image pixels. Independent component analysis (ICA) is a learning rule that produces filters with outputs that are as independent as possible, and is sensitive to high-order dependencies as well as the covariances in the data.

Section 4 also examines an approach in which the image filters were learned from supervised learning. Fishers linear discriminants (FLD) is a supervised learning technique based on the second-order image statistics. It is a linear

projection of the images onto a low dimensional subspace in which the classes are maximally separated.

In Section 5, classification performances with these data-driven image filters are compared to Gabor wavelets, in which the filter kernels are pre-defined, and chosen to model biological vision. Gabor kernels are 2-D sine waves modulated by a Gaussian envelope, and they model receptive fields of simple cells in the primary visual cortex (Daugman, 1988). The representation employed a family of such kernels at 5 spatial frequencies and 8 orientations.

These kernels differ not only in their method of derivation but also in their extent of spatial analysis. Gabor kernels are spatially local, meaning that they analyze a limited portion of the image, whereas kernels such as PCA (eigenfaces) are holistic. Section 5 contrasts holistic and local spatial analysis. Some kernels, such as ICA and LFA, are derived from holistic analysis of the image, but produce local filters in the output. We class these techniques as holistic because of the derivation, but class the resulting filters as local. In order to piece apart local and global filter properties from data-driven versus pre-defined properties, Section 5 also examines local implementations of PCA (Padgett and Cottrell, 1997) in which the kernels were derived from the statistics of small image patches. Similarly, we introduce a multiscale version of the local PCA representation, local PCA jets, to piece apart the multiscale property of the Gabor wavelet representation.

Section 6 provides benchmarks for the performance of the computer vision systems by measuring the ability of naive and expert human subjects to classify the facial actions.

2. IMAGE DATABASE

The system was trained and tested using the database of directed facial actions described in Chapter 5, Section 1. For this investigation, we used data from 20 subjects and attempted to classify 12 actions: 6 upper face actions and 6 lower face actions. See Figure 6.1 for a summary of the actions examined. There were a total of 111 action sequences, (9, 10, 18, 20, 5, 18) respectively of the six upper face actions, and (8, 4, 4, 5, 4, 6) of the six lower face actions. The actions were divided into upper and lower-face categories because facial actions in the lower face have little influence on facial motion in the upper face, and vice versa (Ekman and Friesen, 1978) which allowed us to treat them separately.

The face was located in the first frame in each sequence using the centers of the eyes and mouth. These coordinates were obtained manually by a mouse click. Accurate image registration is critical to holistic approaches such as principal component analysis. An alignment procedure similar to this one was found to give the most accurate image registration during the FERET test (Phillips et al., 1997). The variance in the assigned feature location using this

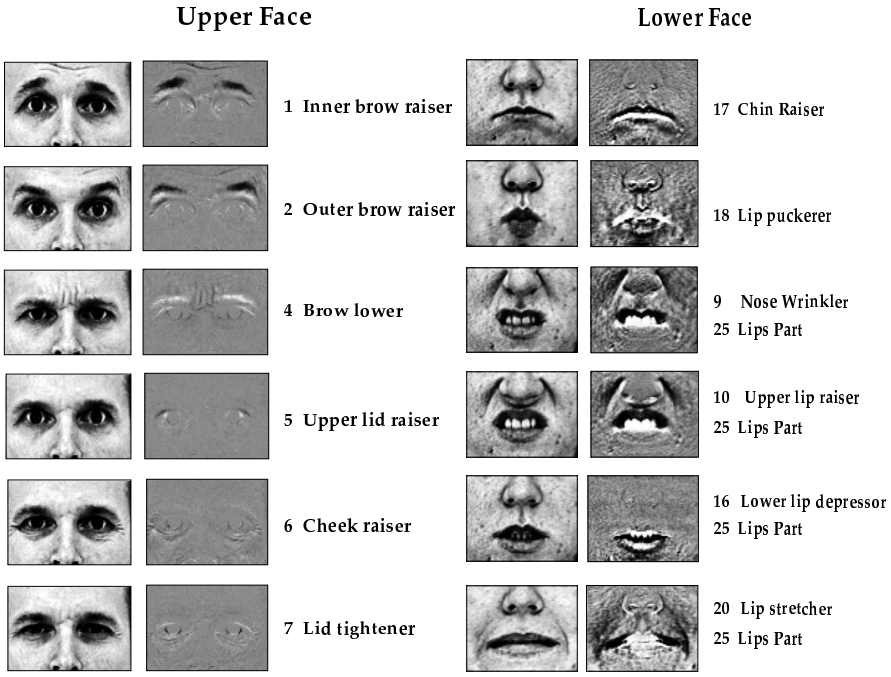

Figure 6.1. List of facial actions classified in this study. From left to right: Example cropped image of the highest magnitude action, the δ image obtained by subtracting the neutral frame (the first image in the sequence), Action Unit number, and Action Unit name.

procedure was 0.4 pixels in the 640×480 pixel images. The coordinates from Frame 1 were used to register the subsequent frames in the sequence. We found in pilot investigations that rigid head motion was smaller than the positional noise in the registration procedure. The three coordinates were used to align the faces, rotate the eyes to horizontal, scale, and finally crop a window of 60×90 pixels containing the region of interest (upper or lower face). The aspect ratios of the faces were warped so that the eye and mouth centers coincided across all images. It has been found that identity recognition performance using principal component based approaches is most successful when the images are warped to remove variations in facial shape (Beymer and Poggio, 1996; Vetter and Troje, 1997).

To control the variation in lighting between frames of the same sequence and in different sequences, we applied a logistic filter with parameters chosen to match the statistics of the grayscale levels of each sequence (Movellan, 1995). This procedure enhanced the contrast, performing a partial histogram equalization on the images.

3. OPTIC FLOW ANALYSIS

The majority of work on facial expression recognition has focused on facial motion analysis through optic flow estimation (Mase, 1991; Essa and Pentland, 1997; Yacoob and Davis, 1994; Rosenblum et al., 1996; Cohn et al., 1999; Lien et al., 2000). While motion is an important cue for facial expression recognition, it is not the only cue. Here we directly compare performance of a motion-based system to that of methods that employ graylevel texture filters. Optic flow fields were estimated by employing a correlation-based technique developed by Singh (Singh, 1991). This algorithm produces flow fields with sub-pixel accuracy, and is comprised of two main components: 1) Local velocity extraction using luminance conservation constraints, 2) Local smoothing.

3.1. Local velocity extraction

We start with a sequence of three images at time $t = t_0 - 1, t_0, t_0 + 1$ and use it to recover all the velocity information available locally. For each pixel $\mathcal{P}(x, y)$ in the central image ($t = t_0$), 1) A small window $\mathcal{W}_p$ of 3×3 pixels is formed around $\mathcal{P}$. 2) A search area $\mathcal{W}_s$ of 5×5 pixels is considered around location (x, y) in the other two images. 3) The correlation between $\mathcal{W}_p$ and the corresponding window centered on each pixel in $\mathcal{W}_s$ is computed, thus giving the matching strength, or *response*, at each pixel in the search window $\mathcal{W}_s$.

At the end of this process, a response distribution $\mathcal{R}$ is defined as the response at each location in $\mathcal{W}_s$. The response at each point gives the frequency of occurrence, or likelihood, of the corresponding value of velocity. (A high response at a location in $\mathcal{W}_s$ that is above and to the right of (x, y) indicates a high likelihood of motion upward and to the right.) Employing a constant temporal model, the response distributions for the two windows corresponding to $t_0 - 1$ and $t_0 + 1$, ($\mathcal{R}_{-1}$ and $\mathcal{R}_{+1}$), are combined by $R = \mathcal{R}_{+1} + \pi\mathcal{R}_{-1}$. Velocity is then estimated using the weighted least squares estimate in (6.1). Figure 6.2 shows an example flow field obtained by this algorithm.

$$\hat{u} = \frac{\sum_u \sum_v \mathcal{R}(u,v)u}{\sum_u \sum_v \mathcal{R}(u,v)} \quad \hat{v} = \frac{\sum_u \sum_v \mathcal{R}(u,v)v}{\sum_u \sum_v \mathcal{R}(u,v)} \quad u, v \in [-2, 2] \tag{6.1}$$

3.2. Local smoothing

To refine the conservation constraint estimate $\mathcal{U}_{cc}$=$(\hat{u}, \hat{v})$ obtained above, a local neighborhood estimate of velocity, $\overline{\mathcal{U}}$, is defined as a weighted sum of the velocities in a neighborhood of $\mathcal{P}$ using a 5×5 Gaussian mask. An optimal estimate $\mathcal{U}$ of (u, v) should combine the two estimates $\mathcal{U}_{cc}$ and $\overline{\mathcal{U}}$, from the conservation and local smoothness constraints respectively. Since $\mathcal{U}$ is a point in (u, v) space, its distance from $\overline{\mathcal{U}}$, weighted by its covariance matrix $\overline{\mathcal{S}}$, represents the error in the smoothness constraint estimate. Similarly,

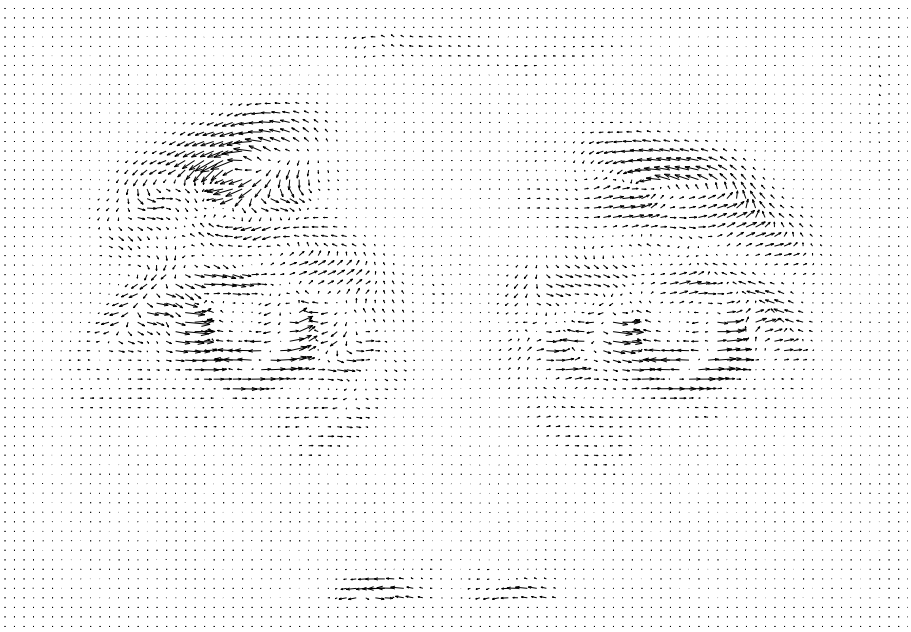

Figure 6.2. Optic flow for AU1 calculated from local velocity information extracted by the correlation-based technique, with no spatial smoothing.

the distance between $\mathcal{U}$ and $\mathcal{U}_{cc}$ weighted by $\mathcal{S}_{cc}$ represents the error due to conservation constraints. Computing $\mathcal{U}$ then, amounts to simultaneously minimizing the two errors:

$$\mathcal{U} = \arg\min\{\|\mathcal{U} - \mathcal{U}_{cc}\|_{\mathcal{S}_{cc}} \quad \bigwedge \quad \|\mathcal{U} - \overline{\mathcal{U}}\|_{\overline{\mathcal{S}}}\}. \tag{6.2}$$

Since we do not know the *true* velocity, this estimate must be computed iteratively. To update the field we use the equations (Singh, 1991):

$$\mathcal{U}^0 = \mathcal{U}_{cc}$$

$$\mathcal{U}^{k+1} = [\mathcal{S}_{cc}^{-1} + \overline{\mathcal{S}}^{-1}]^{-1}[\mathcal{S}_{cc}^{-1}\mathcal{U}_{cc} + \overline{\mathcal{S}}^{-1}\overline{\mathcal{U}}^k] \tag{6.3}$$

where $\overline{\mathcal{U}}^k$ is the estimate derived from smoothness constraints at step k. The iterations stop when

$$\|\mathcal{U}^{k+1} - \mathcal{U}^k\| < \varepsilon$$

with $\varepsilon \propto 10^{-4}$.

3.3. Classification procedure

The following classification procedures were used to test the efficacy of each representation in this comparison for facial action recognition. Each

image analysis algorithm produced a feature vector, f. We employed a simple nearest neighbor classifier in which the similarity S of a training feature vector, f^t, and a novel feature vector, f^n, was measured as the cosine of the angle between them:

$$S(f^n, f^t) = \frac{\langle f^n, f^t \rangle}{\|f^n\| \cdot \|f^t\|} \in [-1, 1]. \tag{6.4}$$

Classification performances were also evaluated using Euclidean distance instead of cosine as the similarity measure and template matching instead of nearest neighbor as the classifier, where the templates consisted of the mean feature vector for the training images. The similarity measure and classifier that gave the best performance is indicated for each technique.

The algorithms were trained and tested using leave-one-out cross-validation, also known as the jack-knife procedure, which makes maximal use of the available data for training. In this procedure, the image representations were calculated multiple times, each time using images from all but one subject for training, and reserving one subject for testing. This procedure was repeated for each of the 20 subjects, and mean classification accuracy was calculated across all of the test cases.

Table 6.1 presents classification performances for the medium magnitude facial actions, which occur in the middle of each sequence. Performance was consistently highest for the medium magnitude actions. Flow fields were calculated from frames 2, 3, and 4 of the image sequence, and the performance of the brightness-based algorithms are presented for frame 4 of each sequence. A class assignment is considered "correct" if it is consistent with the labels assigned by human experts during image collection. The consistency of human experts with each other on this image set is indicated by the agreement rates also shown in Table 6.1.

Optic flow performance

Best performance for the optic flow approach was obtained using the the cosine similarity measure and template matching classifier. The correlation-based flow algorithm gave 85.6% correct classification performance. Since optic flow is a noisy measure, many flow-based expression analysis systems employ regularization procedures such as smoothing and quantizing. We found that spatial smoothing did not improve performance, and instead degraded it to 53.1%. It appears that high spatial resolution optic flow is important for facial action classification. In addition, the motion in facial expression sequences is *nonrigid* and can be highly discontinuous due to the formation of wrinkles. Smoothing algorithms that are not sensitive to these boundaries can be disadvantageous.

There are a variety of choices of flow algorithms, of which Singh's correlation-based algorithm is just one. Also, it is possible that adding more data to the flow

field estimate could improve performance. The results obtained here, however, were comparable to the performance of other facial expression recognition systems based on optic flow (Yacoob and Davis, 1994; Rosenblum et al., 1996). Optic flow estimates can also be further refined, such as with a Kalman filter in an estimation-and control framework, e.g. (Essa and Pentland, 1997). The comparison here addresses direct, image-based representations that do not incorporate a physical model. Sequences of flow fields can also be analyzed using dynamical models such as HMMs or radial basis functions, e.g. (Rosenblum et al., 1996). Such dynamical models could also be employed with texture-based representations. Here we compare all representations using the same classifiers.

4. HOLISTIC ANALYSIS

A number of approaches to face image analysis employ data-driven kernels learned from the statistics of the face image ensemble. Here we explore representations derived from four families of such kernels: eigenfaces, local feature analysis (LFA), Fisher's linear discriminants (FLD), and independent component analysis (ICA).

The holistic spatial analysis algorithms examined in this section each found a set of n-dimensional data-driven image kernels, where n is the number of pixels in each image. The analysis was performed on the difference (or δ) images (Figure 6.1), obtained by subtracting the first image in a sequence (neutral frame) from all of the subsequent frames in each sequence. Advantages of difference images include robustness to changes in illumination, removal of surface variations between subjects, and emphasis of the dynamic aspects of the image sequence (Movellan, 1995). The kernels were derived from low, medium, and high magnitude actions. Holistic kernels for the upper and lower-face subimages were calculated separately.

The methods in this section begin with a data matrix X where the δ-images were stored as row vectors x_j, and the columns had zero mean. In the following descriptions, n is the number of pixels in each image, N is the number of training images and p is the number of principal components retained to build the final representation.

4.1. Principal component analysis: "EigenActions"

Eigenfaces (Turk and Pentland, 1991) employs principal component analysis, which is an unsupervised learning method based on the second-order dependencies among the pixels. Second-order dependencies are pixelwise covariances. Representations based on principal component analysis have been applied successfully to recognizing facial identity (Cottrell and Fleming, 1990; Turk and Pentland, 1991), classifying gender (Cottrell and Metcalfe,

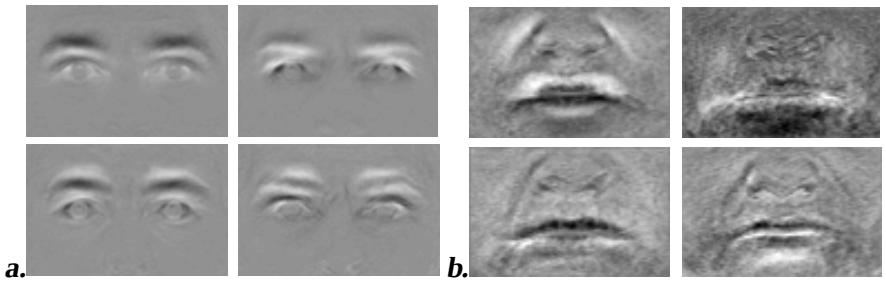

Figure 6.3. First 4 principal components of the difference images for the upper face actions (a), and lower face actions (b). Components are ordered left to right, top to bottom.

1991; Golomb et al., 1991), and recognizing facial expressions (Cottrell and Metcalfe, 1991; Padgett and Cottrell, 1997; Bartlett et al., 1996).

The methods employed here are based on (Cottrell and Metcalfe, 1991) and (Turk and Pentland, 1991), with the primary distinction in that we performed principal component analysis on the dataset of difference images. The principal components were obtained by calculating the eigenvectors of the pixelwise covariance matrix, S, of the δ-images, X. The eigenvectors were found by decomposing S into the orthogonal matrix P and diagonal matrix D: $S = PDP^T$. Examples of the eigenvectors are shown in Figure 6.3. The zero-mean δ-frames of each sequence were then projected onto the first p eigenvectors in P, producing a vector of p coefficients for each image.

Best performance with the holistic principal component representation, 79.3% correct, was obtained with the first 30 principal components, using the Euclidean distance similarity measure and template matching classifier. Previous studies, e.g. (Belhumeur et al., 1997), reported that discarding the first 1 to 3 components improved performance. Here, discarding these components degraded performance.

4.2. Local feature analysis (LFA)

Penev and Atick (Penev and Atick, 1996) recently developed a topographic representation based on second-order image dependencies called local feature analysis (LFA). A representation based on LFA gave the highest performance on the March 1995 FERET face recognition competition (Phillips et al., 1998). The LFA kernels are spatially local, but in this paper we class this technique as holistic, since the image-dimensional kernels are derived from statistical analysis over the whole image.

Local Feature Analysis (LFA) defines a set of topographic, local kernels that are optimally matched to the second-order statistics of the input ensemble (Penev and Atick, 1996). The kernels are derived from the principal component axes, and consist of "sphering" the PCA coefficients to equalize their variance

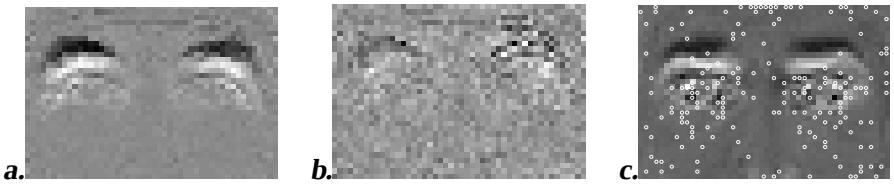

Figure 6.4. a. An original δ-image, b. its corresponding LFA output $O(x)$, and c. the first 155 filter locations selected by the sparsification algorithm superimposed on the mean upper face δ-image.

(Atick and Redlich, 1992), followed by a rotation to pixel space. We begin with the zero-mean matrix of $\delta-$images, X, and calculate the principal component eigenvectors P according to $S = PDP^T$. Penev & Atick (Penev and Atick, 1996) defined a set of kernels, K as

$$K = PVP^T \quad \text{where} \quad V = D^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{\lambda_i}}) \quad i = 1, \ldots, p \tag{6.5}$$

where λ_i are the eigenvalues of S. The rows of K contain the kernels. The kernels were found to have spatially local properties, and are "topographic" in the sense that they are indexed by spatial location (Penev and Atick, 1996). The kernel matrix K transforms X to the LFA output $O = KX^T$ (see Figure 6.4). Note that the matrix V is the inverse square root of the covariance matrix of the principal component coefficients. This transform spheres the principal component coefficients (normalizes their output variance to unity) and minimizes correlations in the LFA output. Another way to interpret the LFA output O is that it is the image reconstruction using sphered PCA coefficients, $O = P(VP^TX^T)$.

4.2.1 Sparsification of LFA

LFA produces an n dimensional representation, where n is the number of pixels in the images. Since we have n outputs described by $p << n$ linearly independent variables, there are residual correlations in the output. Penev & Atick presented an algorithm for reducing the dimensionality of the representation by choosing a subset $\mathcal{M}$ of outputs that were as decorrelated as possible. The sparsification algorithm was an iterative algorithm based on multiple linear regression. At each time step, the output point that was predicted most poorly by multiple linear regression on the points in $\mathcal{M}$ was added to $\mathcal{M}$. Due to the topographic property of the kernels, selection of output points was equivalent to selection of kernels for the representation.

The methods in (Penev and Atick, 1996) addressed image *representation* but did not address *recognition.* The sparsification algorithm in (Penev and Atick, 1996) selected a different set of kernels, $\mathcal{M}$, for each image, which is

problematic for recognition. In order to make the representation amenable to recognition, we selected a single set $\mathcal{M}$ of kernels for all images. At each time step, the kernel corresponding to the pixel with the largest mean reconstruction error *across all images* was added to $\mathcal{M}$.

At each step, the kernel added to $\mathcal{M}$ is chosen as the kernel corresponding to location

$$\arg\max\langle\|O - O^{rec}\|^2\rangle \tag{6.6}$$

where O^{rec} is a reconstruction of the complete output, O, using a linear predictor on the subset of the outputs O generated from the kernels in $\mathcal{M}$. The linear predictor is of the form:

$$\mathcal{Y} = \beta\mathcal{X} \tag{6.7}$$

where $\mathcal{Y} = O^{rec}$, β is the vector of the regression parameters, and $\mathcal{X} = O(\mathcal{M}, N)$. Here $O(\mathcal{M}, N)$ denotes the subset of O corresponding to the points in $\mathcal{M}$ for all N images.[2] β is calculated from:

$$\beta = \frac{\mathcal{Y}\mathcal{X}}{(\mathcal{X}^T\mathcal{X})} = \frac{(O^{rec})^T O(\mathcal{M}, \mathcal{N})}{\mathcal{O}(\mathcal{M}, \mathcal{N})^{\mathcal{T}}\mathcal{O}(\mathcal{M}, \mathcal{N})}. \tag{6.8}$$

Equation 6.8 can also be expressed in terms of the correlation matrix of the outputs, $C = O^T O$, as in (Penev and Atick, 1996):

$$\beta = C(\mathcal{M}, N)C(\mathcal{M}, \mathcal{M})^{-1}. \tag{6.9}$$

The termination condition was $|\mathcal{M}| = N$. Figure 6.4 shows the locations of the points selected by the sparsification algorithm for the upper-face images. We evaluated classification performance using the first i kernels selected by the sparsification algorithm, up to $N = 155$.

The local feature analysis representation attained 81.1% correct classification performance. Best performance was obtained using the first 155 kernels, the cosine similarity measure, and nearest neighbor classifier. Classification performance using LFA was not significantly different from the performance using global PCA. Although a face recognition algorithm related to LFA outperformed eigenfaces in the March 1995 FERET competition (Phillips et al., 1998), our results suggest that an aspect of the algorithm other than the LFA representation accounts for the difference in performance. The exact algorithm used in the FERET test was not disclosed at the time of this research.

[2] $O(\mathcal{M}, N) = O(i, j), \forall i \in \mathcal{M}, \forall j = 1, \ldots, N$.

4.3. *"FisherActions"*

Another holistic image representation that has recently shown to be effective for identity recognition is based on Fisher's Linear discriminants (FLD) (Belhumeur et al., 1997). FLD is a supervised learning method that uses second-order statistics to find a class-specific linear projection of the images. FLD projects the images into a subspace in which the classes are maximally separated (Fisher, 1936). Belhumeur and others (Belhumeur et al., 1997) showed that an FLD projection of a principal components representation of faces improved identity recognition performance. FLD assumes linear separability of the classes. For identity recognition, the approach relied on the assumption that images of the same face under different viewing conditions lie in an approximately linear subspace of the image space, an assumption which holds true for changes in lighting if the face is modeled by a Lambertian surface (Shashua, 1992; Hallinan, 1995). In our dataset, the lighting conditions are fairly constant and most of the variation is suppressed by the logistic filter. The linear assumption for facial expression classification is that the $\delta-$images of a facial action across different faces lie in a linear subspace.

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. Let $\chi \stackrel{\Delta}{=} \{\chi_1, \chi_2, \ldots, \chi_c\}$ be the set of all $N = |\chi|$ data, divided into c classes. Each class χ_i is composed of a variable number of images $x_i \in \mathrm{R}^n$. The between–class scatter matrix S_B and the inter–class scatter S_W are defined as

$$S_B \stackrel{\Delta}{=} \sum_{i=1}^{c} |\chi_i|(\mu_i - \mu)(\mu_i - \mu)^T \text{ and } S_W \stackrel{\Delta}{=} \sum_{i=1}^{c} \sum_{x_k \in \chi_i} (x_k - \mu_i)(x_k - \mu_i)^T \tag{6.10}$$

where μ_i is the mean image of class χ_i and μ is the mean of all data. W_{opt} projects $\mathrm{R}^n \mapsto \mathrm{R}^{c-1}$ and satisfies

$$W_{opt} = \arg\max_W J(W) \stackrel{\Delta}{=} \arg\max_W \frac{\det(W^T S_B W)}{\det(W^T S_W W)} = \{w_1, w_2, \ldots, w_{c-1}.\} \tag{6.11}$$

The $\{w_i\}$ are the solutions to the generalized eigenvalues problem $S_B w_i = \lambda_i S_W w_i$ for $i = 1, \ldots, c-1$. Following (Belhumeur et al., 1997), the calculations are greatly simplified by first performing PCA on the total scatter matrix $S_T = S_W + S_B$ to project the feature space to R^p. Denoting the PCA projection matrix W_{pca}, we project S_W and S_B:

$$\tilde{S}_B \stackrel{\Delta}{=} W_{pca}^T S_B W_{pca} \quad \text{and} \quad \tilde{S}_W \stackrel{\Delta}{=} W_{pca}^T S_W W_{pca}. \tag{6.12}$$

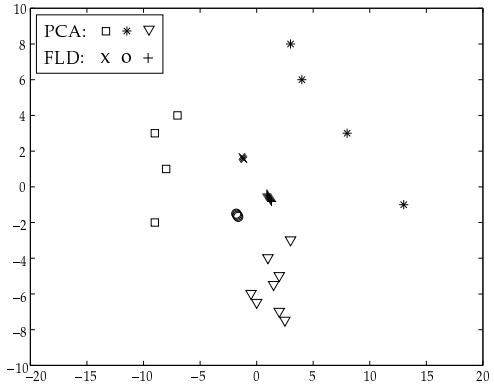

Figure 6.5. PCA and FLD projections of three lower-face action classes onto two dimensions. FLD projections are slightly offset for visibility. FLD projected each class to a single point.

The original FLD problem is thus reformulated as:

$$W_{fld} = \arg\max_W J(W) \triangleq \arg\max_W \frac{\det(W^T \tilde{S}_B W)}{\det(W^T \tilde{S}_W W)} = \{w'_1, w'_2, \ldots, w'_{c-1}\}. \tag{6.13}$$

From 6.11 and 6.13, $W_{opt} = W_{pca} W_{fld}$, and the $\{w'_i\}$ can now be calculated using $\tilde{S}_W^{-1} \tilde{S}_B w'_i = \lambda_i w'_i$ where $\tilde{S}_W$ is full-rank for $p \leq N - c$.

Best performance was obtained by choosing $p = 30$ principal components to first reduce the dimensionality of the data. The data was then projected down to 5 dimensions via the projection matrix, W_{fld}. Best performance of 75.7% correct was obtained with the Euclidean distance similarity measure and template matching classifier.

Clustering with FLD is compared to PCA in Figure 6.5. As an example, three lower face actions were projected down to $c - 1 = 2$ dimensions using FLD and PCA. The FLD projection virtually eliminated within-class scatter of the training set, and the the exemplars of each class were projected to a single point. The three actions in this example were 17, 18, and 9+25.

Contrary to the results obtained in (Belhumeur et al., 1997), Fisher's Linear Discriminants did not improve classification over basic PCA (eigenfaces), despite providing a much more compact representation of the data that optimized linear discrimination. This suggests that the linear subspace assumption was violated more catastrophically for our dataset than for the dataset in (Belhumeur et al., 1997) which consisted of faces under different lighting conditions. Another reason for the difference in performance may be due to the problem of generalization to novel subjects. The FLD method achieved the best performance on the training data (close to 100%) but generalized poorly to new individuals. This is consistent with other reports of poor generalization to novel

subjects (Chellappa, 1998) (also H. Wechsler, personal communication). Good performance with FLD has only been obtained when other images of the test subject were included in the training set. The low dimensionality may provide insufficient degrees of freedom for linear discrimination between classes of face images (Chellappa, 1998). Class discriminations that are approximately linear in high dimensions may not be linear when projected down to as few as 5 dimensions.

4.4. Independent component analysis

Representations such as eigenfaces, LFA, and FLD are based on the second-order dependencies of the image set, the pixelwise covariances, but are insensitive to the high-order dependencies of the image set. High-order dependencies in an image include nonlinear relationships among the pixel grayvalues such as edges, in which there is phase alignment across multiple spatial scales, and elements of shape and curvature. In a task such as facial expression analysis, much of the relevant information may be contained in the high-order relationships among the image pixels. Independent component analysis (ICA) is a generalization of PCA which learns the high-order moments of the data in addition to the second-order moments. Chapter 3 developed face representations using ICA which are based on the high-order in addition to the second-order dependencies in the images (Bartlett and Sejnowski, 1997; Bartlett et al., 1998; Bartlett, 1998). In a direct comparison, the ICA representations outperformed PCA for identity recognition. The methods in this section employ Architecture I, described in Chapter 3, Section 2.

The independent component representation was obtained by performing "blind separation" on the set of face images (Bartlett and Sejnowski, 1997; Bartlett et al., 1998; Bartlett, 1998). In the image synthesis model of Figure 6.6, the δ – images in the rows of X are assumed to be a linear mixture of an unknown set of statistically independent source images S, where A is an unknown mixing matrix. The sources are recovered by a learned unmixing matrix W, which approximates A^{-1} and produces statistically independent outputs, U.

The ICA unmixing matrix W was found using an unsupervised learning algorithm derived from the principle of optimal information transfer between neurons (Bell and Sejnowski, 1995; Bell and Sejnowski, 1997). The algorithm maximizes the mutual information between the input and the output of a nonlinear transfer function g. A discussion of how information maximization leads to independent outputs can be found in (Nadal and Parga, 1994; Bell and Sejnowski, 1995; Bell and Sejnowski, 1997). Let $u = Wx$ where x is a column of the image matrix X, and $y = g(u)$. The update rule for the weight matrix, W, is given by

$$\Delta W = (I + y'u^T)W \tag{6.14}$$

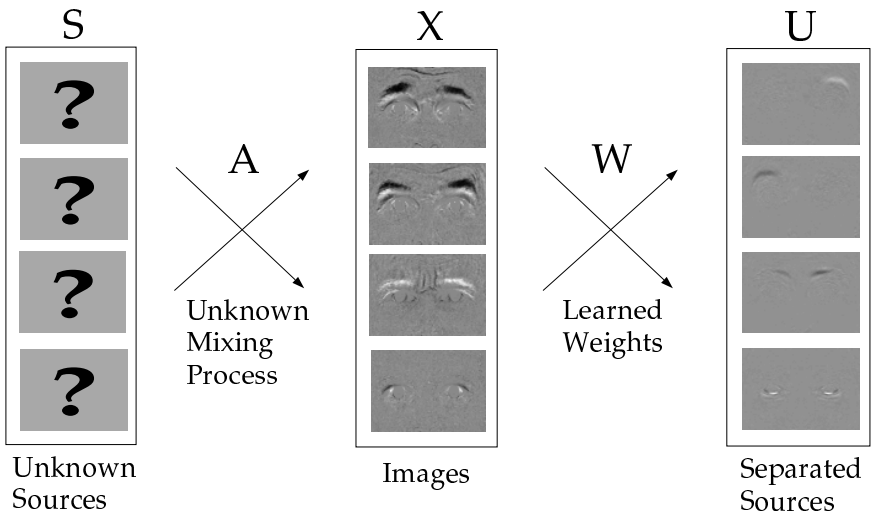

Figure 6.6. Image synthesis model for the ICA representation.

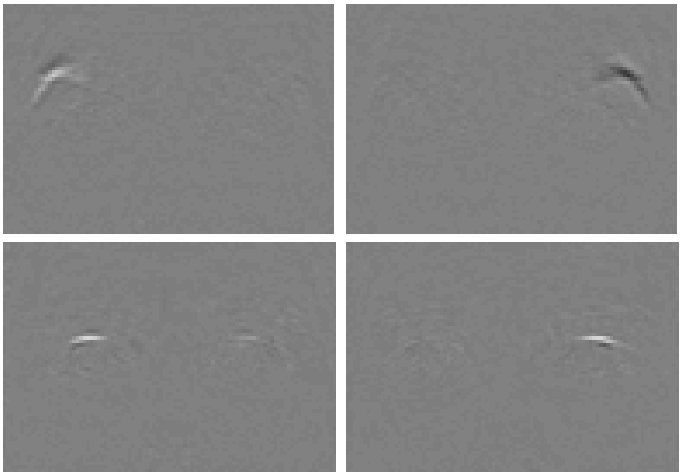

Figure 6.7. Sample ICA basis images.

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

We employed the logistic transfer function, $g(u) = \frac{1}{1+e^{-u}}$, giving $y' = (1 - 2y_i)$. Convergence is greatly speeded by including a "sphering" step prior to learning (Bell and Sejnowski, 1997), in which the zero-mean dataset X is passed through the whitening filter, $W_Z = 2 * \langle XX^T \rangle^{-\frac{1}{2}}$. This removes both the first and the second-order dependencies from the data. The full transform

was therefore $W = W_I * W_Z$, where W_I is the weight obtained by information maximization in Equation 6.14.

The projection of the image set onto each weight vector in W produced an image of the statistical dependencies that each weight vector learned. These images are the rows of the output matrix U, and examples are shown in Figure 6.7. The rows of U are the independent components of the image set, and they provided a basis set for the expression images. The ICA representation consisted of the coefficients, a, for the linear combination of basis images in U that comprised each face image in X. These coefficients were obtained from the rows of the estimated mixing matrix $A \stackrel{\Delta}{=} W^{-1}$ (Bartlett et al., 1998). The number of independent components extracted by the ICA algorithm corresponds with the number of input images. Two hundred independent components were extracted for the upper and 155 for the lower face image sets. Since there were more than 200 upper face images, ICA was performed on 200 linear mixtures of the faces without affecting the image synthesis model. The first 200 PCA eigenvectors were chosen for these linear mixtures since they give the combination of images that accounts for the maximum variability among the pixels. The eigenvectors were normalized to unit length. Details are available in Chapter 3

Unlike PCA, there is no inherent ordering to the independent components of the dataset. We therefore selected as an ordering parameter the class discriminability of each component. Let $\overline{a}_k$ be the overall mean of coefficient a_k, and $\overline{a}_{jk}$ be the mean for action j. The ratio of between-class to within-class variability, r, for each coefficient is defined as

$$r = \frac{\sigma_{between}}{\sigma_{within}} \tag{6.15}$$

where $\sigma_{between} = \sum_j (\overline{a}_{jk} - \overline{a}_k)^2$ is the variance of the j class means, and $\sigma_{within} = \sum_j \sum_i (a_{ijk} - \overline{a}_{jk})^2$ is the sum of the variances within each class. The first p components selected by class discriminability comprised the independent component representation.

Best performance of 95.5% was obtained with the first 75 components selected by class discriminability, using the cosine similarity measure, and nearest neighbor classifier. Independent component analysis gave the best performance among all of the holistic classifiers. Note, however, that the independent component images in Figure 6.7 were local in nature. As in LFA, the ICA algorithm analyzed the images as whole, but the basis images that the algorithm learned were local. Two factors contributed to the local property of the ICA basis images: Most of the statistical dependencies were in spatially proximal image locations, and secondly, the ICA algorithm produces sparse outputs (Bell and Sejnowski, 1997).

5. LOCAL REPRESENTATIONS

In the approaches described in Section 4, the kernels for the representation were learned from the statistics of the entire image. Some of these analyses produced global filters, others produced local filters, filters that act on small spatial regions within the image. A number of researchers have argued that local filters are superior to global ones for face image analysis, and expression analysis in particular (Penev and Atick, 1996; Padgett and Cottrell, 1997; Gray et al., 1997; Zhang et al., 1997; Lee and Seung, 1999). Section 5 explores local representations in which the filters are derived from local spatial analysis and/or explicitly act on small spatial regions within the images. We examine three variations on local filters that employ PCA, and compare them to the biologically inspired Gabor wavelet decomposition.

A simple benchmark for the local filters consisted of a single Gaussian kernel. The δ – images were convolved with a 15×15 Gaussian kernel and the output was downsampled by a factor of 4. The dimensionality of the final representation was $\frac{n}{4}$. The output of this basic local filter was classified at 70.3% accuracy using the Euclidean distance similarity measure and template matching classifier.

5.1. Local PCA

There is evidence from a number of sources that local spatial filters may be superior to global spatial filters for facial expression classification. Padgett & Cottrell (Padgett and Cottrell, 1997) found that "eigenfeatures", consisting of the principal components of image subregions containing the mouth and eyes, were more effective than global PCA (full-face eigenfaces) for facial expression recognition. Furthermore, they found that a set of shift-invariant local basis functions derived from the principal components of small image patches were more effective than both eigenfeatures and global PCA. This finding is supported by Gray, Movellan & Sejnowski (Gray et al., 1997) who found that a similar local PCA representation gave better performance than global PCA for lipreading from video.

The methods employed here are based on (Padgett and Cottrell, 1997). The shift-invariant local basis functions employed in (Padgett and Cottrell, 1997) were derived from the principal components of small image patches from randomly sampled locations in the face image. Principal component analysis of image patches sampled from random locations, such that the image statistics are stationary over the patch, describes the amplitude spectrum (Field, 1994; Pratt, 1978).

A set of more than 7000 patches of size 15×15 was taken from random locations in the δ – images and decomposed using PCA. The first p principal components were then used as convolution kernels to filter the full images. The

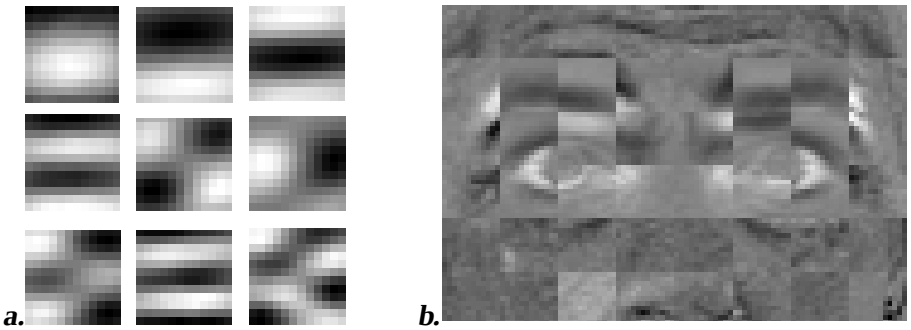

Figure 6.8. a. Shift-invariant local PCA kernels. First 9 components, ordered left to right, top to bottom. b. Shift-variant local PCA kernels. The first principal component is shown for each image location.

outputs were subsequently downsampled by a factor of 4, such that the final dimensionality of the representation was isomorphic to $R^{p \times n/4}$. The local PCA filters obtained from the set of lower-face δ-images are shown in Figure 6.8.

Performance improved by excluding the first principal component. Best performance of 73.4% was obtained with principal components 2-30, using Euclidean distance and template matching. Unlike the findings in (Padgett and Cottrell, 1997), shift invariant basis functions obtained through local PCA were no more effective than global PCA for facial action coding. Performance of this local PCA technique was not significantly higher than that obtained using a single 15x15 Gaussian kernel.

Because the local PCA implementation differed from global PCA in two properties, spatial locality and image alignment, we repeated the local PCA analysis at fixed spatial locations. PCA of location-independent images captures amplitude information without phase, whereas alignment of the images provides implicit phase information (Field, 1994; Bell and Sejnowski, 1997). Local PCA at fixed image locations is related to the eigenfeatures representation addressed in (Padgett and Cottrell, 1997). The eigenfeature representation in (Padgett and Cottrell, 1997) differed from shift-invariant local PCA in image patch size. Here, we compare shift-invariant and shift-variant versions of local PCA while controlling for patch size.

The images were divided into $m \ll \frac{n}{4}$ 15×15 fixed regions. The principal components of each region were calculated separately. Each image was thus represented by $p \times m$ coefficients. The final representation consisted of $p = 10$ principal components of $m = 48$ image regions.

Classification performance was tested using up to the first 30 components of each patch. Best performance of 78.3% was obtained with the first 10 principal components of each image patch, using Euclidean distance and the nearest neighbor classifier. There is a trend for phase alignment to improve

classification performance using local PCA, but the difference is not statistically significant. Contrary to the findings in (Padgett and Cottrell, 1997) neither local PCA representation outperformed the global PCA representation. It has been proposed that local representations reduce sensitivity to identity-specific aspects of the face image (Padgett and Cottrell, 1997; Gray et al., 1997). The success of global PCA here could be attributable to the use of δ images, which reduced variance related to identity specific aspects of the face image. Another reason for the difference in findings could be the method of downsampling. Padgett and Cottrell selected filter outputs from 7 image locations at the eyes and mouth, whereas here downsampling was performed in a grid-wise fashion from 48 image locations.

5.2. Gabor wavelet representation

An alternative to adaptive local filters such as local PCA are pre-defined local filters such as families of Gabor filters. Gabor filters are obtained by modulating a 2-D sine wave with a Gaussian envelope. Such filters remove most of the variability in images due to variation in lighting and contrast, and closely model the response properties of visual cortical cells (Pollen and Ronner, 1981; Jones and Palmer, 1987; DeValois and DeValois, 1988; Daugman, 1988). Representations based on the outputs of families of Gabor filters at multiple spatial scales, orientations, and spatial locations, have proven successful for recognizing facial identity in images (Lades et al., 1993; Phillips et al., 1997). In a direct comparison of face recognition algorithms, Gabor filter representations gave better identity recognition performance than representations based on principal component analysis (Zhang et al., 1997). A Gabor representation was also more effective than a representation based on the geometric locations of facial features for expression recognition (Zhang et al., 1998).

Here we examine a local representation that employs Gabor wavelet decomposition. The methods employed here are based on those described in (Lades et al., 1993). Given an image $\mathcal{I}(\vec{x})$ (where $\vec{x} = (x, y)$), the transform $\mathcal{J}_i$ is defined as a convolution

$$\mathcal{J}_i = \int \mathcal{I}(\vec{x}) \psi_i(\vec{x} - \vec{x}') d^2\vec{x}' \tag{6.16}$$

with a family of Gabor kernels ψ_i

$$\psi_i(\vec{x}) = \frac{||\vec{k}_i||^2}{\sigma^2} e^{-\frac{||\vec{k}_i||^2 ||\vec{x}||^2}{2\sigma^2}} \left[e^{j\vec{k}_i\vec{x}} - e^{-\frac{\sigma^2}{2}} \right]. \tag{6.17}$$

Each ψ_i is a plane wave characterized by the vector $\vec{k}_i$ enveloped by a Gaussian function, where the parameter $\sigma = 2\pi$ determines the ratio of window width to wavelength. The first term in the square brackets determines the oscillatory

part of the kernel, and the second term compensates for the DC value of the kernel (Lades et al., 1993). The vector $\vec{k_i}$ is defined as

$$\vec{k_i} = \begin{pmatrix} f_\nu \cos \varphi_\mu \\ f_\nu \sin \varphi_\mu \end{pmatrix} \tag{6.18}$$

where

$$f_\nu = 2^{-\frac{\nu+2}{2}}\pi, \quad and \quad \varphi_\mu = \mu\frac{\pi}{8}.$$

The parameters ν and μ define the frequency and orientation of the kernels. We used 5 frequencies $(\nu = 0-4)$ and 8 orientations, $(\mu = 1-8)$ in the final representation, following the methods in (Lades et al., 1993). Example filters are shown in Figure 6.9. Spatial frequencies corresponded to 4 - 16 pixels per cycle in 1/2 octave steps. In terms of the face, this corresponds to 2.8 - 12.2 cycles between the eyes, measured pupil to pupil. The Gabor filters were applied to the δ–images, and magnitudes were extracted. The outputs $\{\mathcal{J}_i\}$ of the 40 Gabor filters were downsampled by a factor q to reduce the dimensionality to $40 \times \frac{n}{q}$, and normalized to unit length, which performed a divisive contrast normalization. We tested the performance of the system using $q = 1, 4, 16$ and found that $q = 16$ yielded the best generalization rate. Best performance was obtained with the cosine similarity measure and nearest neighbor classifier.

Classification performance with the Gabor filter representation was 95.5%. This performance was significantly higher than all other approaches in the comparison except independent component analysis, with which it tied. This finding is supported by Zhang, Yan, & Lades (Zhang et al., 1997) who found that face recognition with the Gabor filter representation was superior to that with a holistic principal component based representation.

To determine which frequency ranges contained more information for action classification, we repeated the tests using subsets of high frequencies $(\nu = 0, 1, 2)$, and low frequencies, $(\nu = 2, 3, 4)$. Performance with the high frequency subset was 92.8%, almost the same as for $\nu = 0, \ldots, 4$, whereas performance with the low frequency subset was 83.8%. The finding that the higher spatial frequency bands of the Gabor filter representation contain more information than the lower frequency bands is consistent with our analysis of optic flow, above, in which reduction of the spatial resolution of the optic flow through smoothing had a detrimental effect on classification performance. It appears that high spatial frequencies are important for this task.

5.3. PCA jets

We next investigated whether the multiscale property of the Gabor wavelet representation accounts for the difference in performance obtained using the Gabor representation and the local PCA representation. To test this hypothesis,

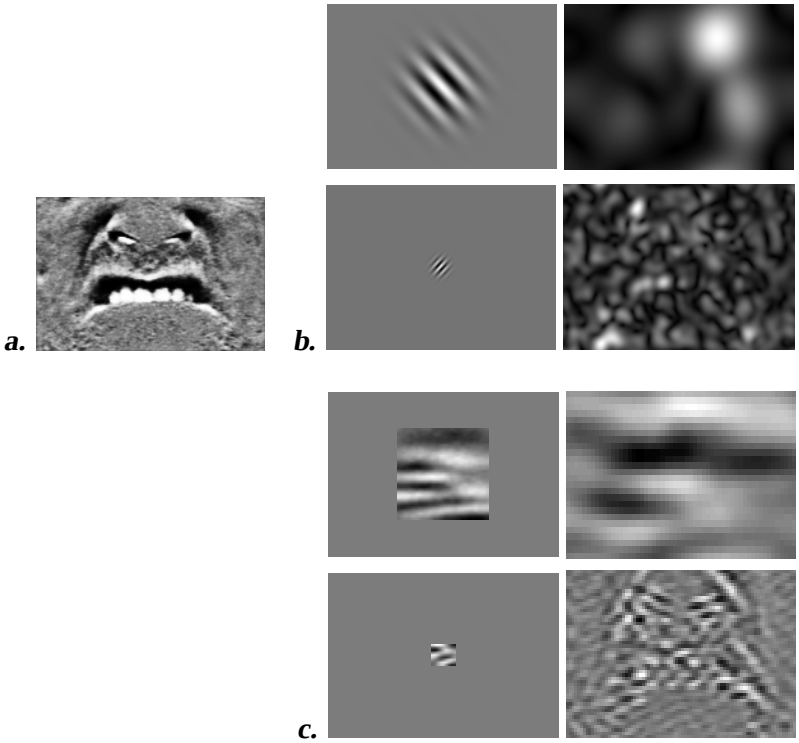

Figure 6.9. a. Original δ-image. b. Gabor kernels (low and high frequency) with the magnitude of the filtered image to the right. c. Local PCA kernels (large and small scale) with the corresponding filtered image.

we developed a multiscale version of the local PCA representation, PCA jets. The principal components of random subimage patches provide the amplitude spectrum of local image regions. A multiscale local PCA representation was obtained by performing PCA on random image patches at five different scales chosen to match the sizes of the Gaussian envelopes (see Figure 6.9). Patch sizes were chosen as $\pm 3\sigma$, yielding the following set: [9×9, 15×15, 23×23, 35 × 35, and 49 × 49]. The number of filters was matched to the Gabor representation by retaining 16 principal components at each scale, for a total of 80 filters. The downsampling factor $q = 16$ was also chosen to match the Gabor representation.

As for the Gabor representation, performance was tested using the cosine similarity measure and nearest neighbor classifier. Best results were obtained using eigenvectors 2 to 17 for each patch size. Performance was 64.9% for all five scales, 72.1% for the three smaller scales, and 62.2% for the three larger scales. The multiscale principal component analysis (PCA jets) did not improve performance over the single scale local PCA. It appears that the

multiscale property of the Gabor representation alone does not account for the improvement in performance obtained with this representation over local representations based on principal component analysis.

6. HUMAN SUBJECTS

The performance of human subjects provided benchmarks for the performances of the automated systems. Most other computer vision systems test performance on prototypical expressions of emotion, which naive human subjects can classify with over 90% agreement, e.g. (McKelvie, 1995). Facial action coding is a more detailed analysis of facial behavior than discriminating prototypical expressions. The ability of naive human subjects to classify the facial action images in this set indicates the difficulty of the visual classification task, and provides a basis for comparing the results presented here with other systems in the literature. Since the long-term goal of this project is to replace human expert coders with with an automated system, a second benchmark was provided by the agreement rates of expert human coders on these images. This benchmark indicated the extent to which the automated systems attained the goal of reaching the consistency levels of the expert coders.

Naive subjects. Naive subjects were ten adult volunteers with no prior knowledge of facial expression measurement. The upper and lower face actions were tested separately. Subjects were provided with a guide sheet which contained an example image of each of the six upper or lower face actions along with a written description of each action and a list of image cues for detecting and discriminating the actions from (Ekman and Friesen, 1978). Each subject was given a training session in which the facial actions were described and demonstrated, and the image cues listed on the guide sheet were reviewed and indicated on the example images. The subjects kept the guide sheet as a reference during the task.

Face images were preprocessed identically to how they had been for the automated systems, as described in Section 2, and printed using a high resolution HP Laserjet 4si printer with 600 dpi. Face images were presented in pairs, with a neutral expression image and the test image presented side by side. Subjects were instructed to compare the test image with the neutral image and decide which of the actions the subject had performed in the test image. Ninety-three image pairs were presented in both the upper and lower face tasks. Subjects were instructed to take as much time as they needed to perform the task, which ranged from 30 minutes to one hour. Naive subjects classified these images at 77.9% correct. Presenting uncropped face images did not improve performance.

Expert coders. Expert subjects were four certified FACS coders. The task was identical to the naive subject task with the following exceptions: Expert subjects were not given a guide sheet or additional training, and the complete

Table 6.1. Best performance for each classifier. PCA: Principal component analysis. LFA: Local feature analysis. FLD: Fisher's linear discriminant. ICA: Independent component analysis. Shift-inv: Shift-invariant. Shift-var: Shift-variant.

Optic Flow	Correlation	85.6% ± 3.3
	Smoothed	53.1% ± 4.7
Holistic Spatial Analysis	PCA	79.3% ± 3.9
	LFA	81.1% ± 3.7
	FLD	75.7% ± 4.1
	ICA	95.5% ± 2.0
Local Spatial Analysis	Gaussian Kernel	70.3 ± 4.
	PCA Shift-inv	73.4% ± 4.2
	PCA Shift-var	78.3% ± 3.9
	PCA Jets	72.1% ± 4.2
	Gabor Jets	95.5% ± 2.0
Human Subjects	Naive	77.9% ± 2.5
	Expert	94.1% ±2.1

face was visible, as it would normally be during FACS scoring. Although the complete action was visible in the cropped images, the experts were experienced with full face images, and the cropping may bias their performance by removing contextual information. One hundred and fourteen upper-face image pairs and ninety-three lower-face image pairs were presented. Time to complete the task ranged from 20 minutes to 1 hour and 15 minutes. The rate of agreement of the expert coders with the assigned labels was was 94.1%.

7. DISCUSSION

We have compared a number of different image analysis methods on a difficult classification problem, the classification of facial actions. Several approaches to facial expression analysis have been presented in the literature, but until now, there has been little direct comparison of these methods on a single dataset. These approaches include analysis of facial motion (Mase, 1991; Yacoob and Davis, 1994; Rosenblum et al., 1996; Essa and Pentland, 1997), holistic spatial pattern analysis using techniques based on principal component analysis (Cottrell and Metcalfe, 1991; Padgett and Cottrell, 1997; Lanitis et al., 1997), and measurements of the shapes and facial features and their spatial arrangements (Lanitis et al., 1997; Zhang et al., 1998). This

investigation compared facial action classification using optic flow, holistic spatial analysis, and local spatial representations. We also included in our comparison a number of representations that had been developed for facial identity recognition, and applied them for the first time to facial expression analysis. These representations included Gabor filters (Lades et al., 1993), Linear Discriminant Analysis (Belhumeur et al., 1997), Local Feature Analysis (Penev and Atick, 1996), and Independent Component Analysis (Bartlett et al., 1998).

Results are summarized in Table 6.1. Best performances were obtained with the local Gabor filter representation, and the Independent Component representation, which both achieved 96% correct classification. The performance of these two methods equaled the agreement level of expert human subjects on these images. Image representations derived from the second-order statistics of the dataset (PCA and LFA) performed about as well as *naive* human subjects on this image classification task, in the 80% accuracy range. Performances using LFA and FLD did not significantly differ from PCA, nor did spatially local implementations of PCA. Correlation-based optic flow performed at a level between naive and expert human subjects, at 86%. Classification accuracies obtained here compared favorably with other systems developed for emotion classification, despite the additional challenges of classifying facial actions over classifying prototypical expressions reviewed in (Hager and Ekman, 1995).

A number of the representations explored here employed spatial filters on the image graylevels. The filter kernels are juxtaposed in Figure 6.10. We obtained converging evidence that local spatial filters are important for analysis of facial expressions. The two representations that significantly outperformed the others, the Gabor representation (Lades et al., 1993) and the Independent Component representation (Bartlett et al., 1998), each employed local filters. ICA was classified as a holistic algorithm since the analysis was performed over the images as a whole, however the basis images that the algorithm produced were local. This supports recent findings that local filters are important for face image analysis (Padgett and Cottrell, 1997; Gray et al., 1997; Lee and Seung, 1999). Our results also demonstrated that spatial locality of the image filters alone is insufficient for good classification. Local principal component representations such as LFA and local PCA performed no better than the global PCA representation (Eigenfaces).

We also obtained multiple sources of evidence that high spatial frequencies are important for classifying facial actions. Spatial smoothing of optic flow degraded performance by more than 30%. Secondly, classification with only the high frequencies of the Gabor representation was superior to classification using only the low spatial frequencies. A similar result was obtained with the PCA jets. These findings are in contrast to a recent report that the information for recognizing prototypical facial expressions was carried predominantly by

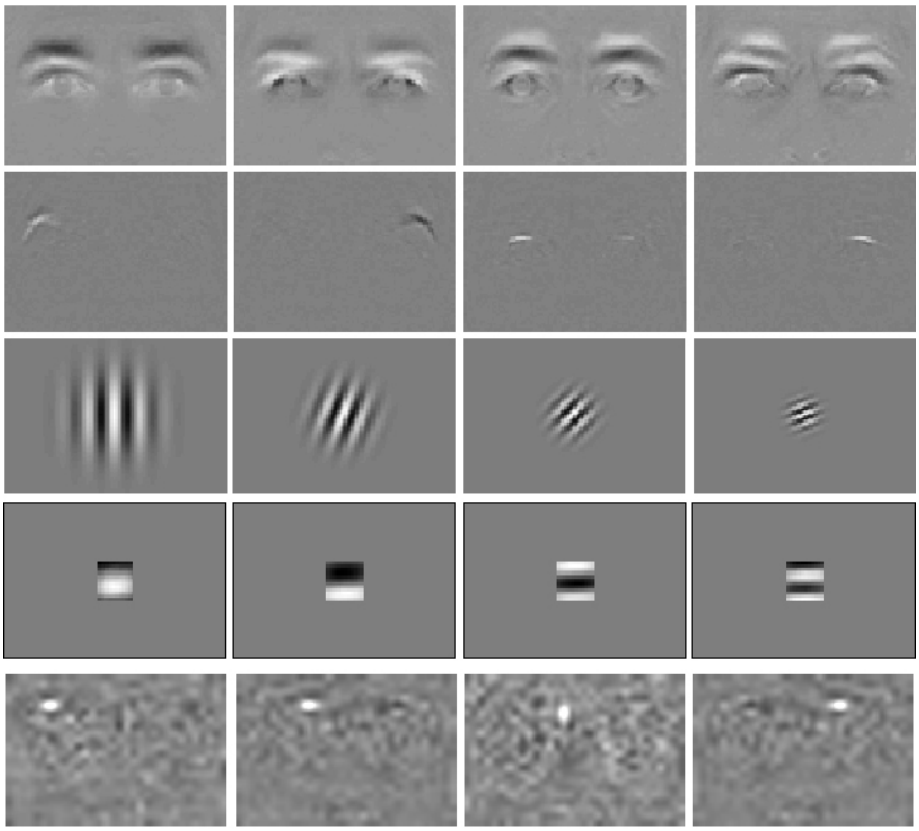

Figure 6.10. Summary of filter kernels. From top to bottom: First 4 PCA kernels; Four ICA kernels. The ICA kernels are local, spatially opponent, and adaptive; Gabor kernels are local, spatially opponent, and predefined; First 4 local PCA kernels; Four LFA kernels.

the low spatial frequencies (Zhang et al., 1998). This difference in findings highlights the difference in the task requirements of classifying facial actions versus classifying prototypical expressions of emotion. Classifying facial actions is a more detailed level of analysis. Our findings predict, for example, that high spatial frequencies would carry important information for discriminating genuine expressions of happiness from posed ones, which differ in the presence of AU 6 (the cheek raiser) (Ekman et al., 1988).

The relevance of high spatial frequencies has implications for motion-based facial expression analysis. Since optic flow is a noisy measure, many flow-based expression analysis systems employ regularization procedures such as smoothing and quantizing to estimate a principal direction of motion within an image region. The analysis presented here suggests that high spatial resolution

optic flow is important for analysis of facial behavior at the level of facial action coding.

In addition to spatial locality, the ICA representation and the Gabor filter representation appear to share other properties including sensitivity to high order dependencies, and relationships to representations in the visual cortex. Bell & Sejnowski (Bell and Sejnowski, 1997) found using ICA that the filters that produced independent outputs from natural scenes were spatially local, oriented edge filters, similar to a bank of Gabor filters, and to the response properties of primary visual cortical cells. It has also been shown that Gabor filter outputs of natural images are sparse and kurtotic (Field, 1987; Field, 1994), and at least pairwise independent[3] (Simoncelli, 1997).

The Gabor wavelets, PCA, and ICA each provide a way to represent face images as a linear superposition of basis functions. Gabor wavelets employ a set of pre-defined basis functions, whereas PCA and ICA learn basis functions that are adapted to the data ensemble. 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 above, there are a number of relationships between Gabor wavelets and the basis functions obtained with ICA. The Gabor wavelets are not specialized to the particular data ensemble, but would be advantageous when the amount of data is too small to estimate filters.

The ICA representation performed as well as the Gabor representation, despite having two orders of magnitude fewer basis functions. A large number of basis functions does not appear to confer an advantage for classification. The PCA-*jet* representation, which was matched to the Gabor representation for number of basis functions as well as scale, performed at only 72% correct.

Each of the local representations underwent downsampling. The effect of downsampling on generalization rate was examined in the Gabor representation, and we found downsampling improved generalization performance. The downsampling was done in a grid-wise fashion, and there was no manual selection of facial features. Downsampling methods can have a significant impact on performance. The Rockefeller group recently revealed that the high performance on the FERET face recognition competition was obtained by employing a different technique for downsampling the LFA representation (Penev, 2001). This involved finding pixel locations with maximum information for

[3]This holds when the responses undergo divisive normalization, which neurophysiologists have proposed takes place in the visual cortex (Heeger, 1991). The length normalization in our Gabor representation is a form of divisive normalization.

each individual. We are presently investigating whether different methods of downsampling could improve performace with Gabors and ICA.

An outstanding issue is whether our findings depend on the simple recognition engines we employed. Would a smarter recognition engine alter the relative performances of the image representations we tested? Our preliminary investigations suggest that is not the case (Bartlett et al., 2000). Hidden Markov Models (HMM's) were trained on the PCA, ICA, and Gabor representations. The Gabor representation was first reduced to 75 dimensions using PCA before training the HMM. The HMM improved performance with ICA to 96.3% correct, and gave similar percent improvements to the Gabor and PCA representations over their nearest neighbor performances.

8. CONCLUSIONS

The results of this comparison provided converging evidence for the importance of using local filters, high spatial frequencies, and statistical independence for classifying facial actions. Best performances were obtained with Gabor wavelet decomposition and independent component analysis. These two representations are related to each other. They employ graylevel texture filters that share properties of spatial locality, independence, and have relationships to the response properties of visual cortical neurons.

The majority of the approaches to facial expression recognition by computer have focused exclusively on analysis of facial motion. Motion is an important aspect of facial expressions, but not the only cue. Although experiments with point-light displays have shown that human subjects *can* recognize facial expressions from motion signals alone (Bassili, 1979), recognition rates are just above chance, and substantially lower than those reported for recognizing a similar set of expressions from static graylevel images, e.g. (McKelvie, 1995). In this comparison, best performances were obtained with representations based on surface graylevels. A future direction of this work is to combine the motion information with spatial texture information. Perhaps combining motion and graylevel information will ultimately provide the best facial expression recognition performance, as it does for the human visual system (Bassili, 1979; Wallbott, 1992).

Acknowledgements

This chapter was based on "Classifying Facial Actions" by G.L. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, which appeared in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(10) p. 974-989, 1999. Most of the research was conducted by G.L. Donato, and most of the writing by M.S. Bartlett. The research was supported by NSF Grant

No. BS-9120868, Lawrence Livermore National Laboratories Intra-University Agreement B291436, Howard Hughes Medical Institute, and NIH Grant No 1 F32 MH12417-01. We are indebted to FACS experts Linda Camras, Wil Irwin, Irene McNee, Harriet Oster, and Erica Rosenberg for their time and assistance with this project. We thank Beatrice Golomb, Wil Irwin, and Jan Larsen for contributions to project initiation, Claudia Hilburn Methvin for image collection, and Laurenz Wiskott and Gary Cottrell for valuable discussions on earlier drafts of this paper.

Chapter 7

LEARNING VIEWPOINT INVARIANT REPRESENTATIONS OF FACES IN AN ATTRACTOR NETWORK

Reprint in full of "Learning viewpoint invariant face representations from visual experience in an attractor network" by Bartlett, M.S. & Sejnowski, T.J., which appeared in *Network: Computation in Neural Systems* 9(3) 399-417, 1998. Reprinted with permission from Institute of Physics Publishing, Ltd. See http://www.iop.org/Journals/nc.

Abstract In natural visual experience, 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. A set of simulations is presented which demonstrate how viewpoint invariant representations of faces can be developed from visual experience by capturing the temporal relationships among the input patterns. The simulations explored the interaction of temporal smoothing of activity signals with Hebbian learning (Földiák, 1991) in both a feedforward layer and a second, recurrent layer of a network. The feedforward connections were trained by Competitive Hebbian Learning with temporal smoothing of the post-synaptic unit activities (Bartlett and Sejnowski, 1996b). The recurrent layer was a generalization of a Hopfield network with a lowpass temporal filter on all unit activities. The combination of basic Hebbian learning with temporal smoothing of unit activities produced an attractor network learning rule that associated temporally proximal input patterns into basins of attraction. These two mechanisms were demonstrated in a model that took graylevel images of faces as input. Following training on image sequences 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.

1. INTRODUCTION

Cells in the primate inferior temporal lobe have been reported that respond selectively to faces despite substantial changes in viewpoint (Perrett et al., 1989; Hasselmo et al., 1989). A small proportion of cells gave responses that were invariant to angle of view, whereas other cells that have been classed as

viewpoint *dependent* had tuning curves that were quite broad. Perrett et al. (Perrett et al., 1989) reported broad coding for five principal views of the head: Frontal, left profile, right profile, looking up, and looking down, and the pose tuning of these cells was on the order of $\pm 40°$. The retinal input changes considerably under these shifts in viewpoint.

This model addresses how receptive fields with such broad pose tuning could be developed from visual experience. The model touches on several issues in the psychology and neurophysiology of face recognition. Can general learning principles account for the ability to respond to faces across changes in pose, or does this function require special purpose, possibly genetically encoded mechanisms? Is it possible to recognize faces across changes in pose without explicitly recovering or storing the 3-dimensional structure of the face? What are the potential contributions of temporal sequence information to the representation and recognition of faces?

Until recently, most investigations of face recognition focused on static images of faces. The preponderance of our experience with faces, however, is not with static faces, but with live faces that move, change expression, and pose. Temporal sequences contain information that can aid in the process of representing and recognizing faces and objects, e.g. (Bruce, 1988). This model explores how a neural system can acquire invariance to viewpoint from visual experience by accessing the temporal structure of the input. The appearance of an object or a face changes continuously as the observer moves through the environment or as a face changes expression or pose. Capturing the temporal relationships in the input is a way to automatically associate different views of an object without requiring three-dimensional representations (Stryker, 1991b).

Temporal association may be an important factor in the development of pose invariant responses in the inferior temporal lobe 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). Macaques were presented a fixed sequence of 97 fractal patterns for 2 weeks. After training, the patterns were presented in random order. Figure 7.1 shows correlations in sustained responses of the AIT cells to pairs of patterns as a function of the relative position of the patterns in the training sequence. Responses to neighboring patterns were correlated, and the correlation dropped off as the distance between the patterns in the training sequence increased. These data suggest that cells in the temporal lobe can modify their receptive fields to associate patterns that occurred close together in time.

Hebbian learning can capture temporal relationships in a feedforward system when the output unit activities undergo temporal smoothing (Földiák, 1991). This mechanism learns viewpoint-tolerant representations when different views

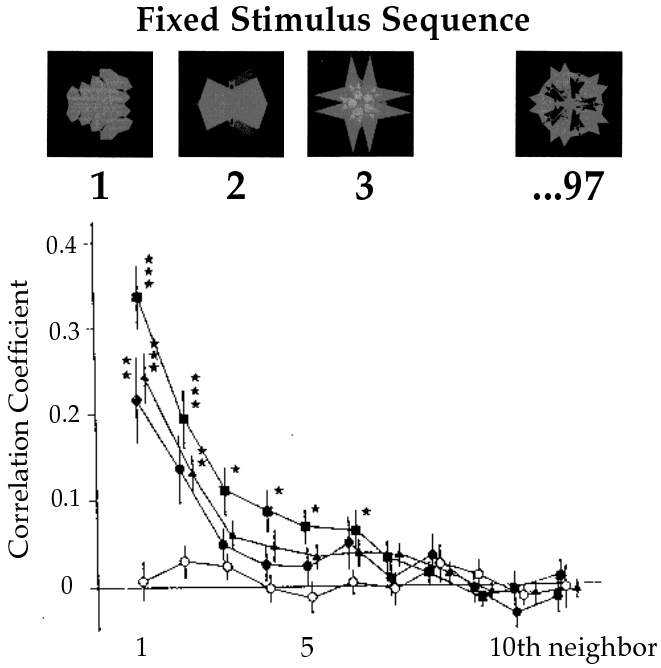

Figure 7.1. Evidence of temporal associations in IT. Top: Samples of the 97 fractal pattern stimuli in the fixed training sequence. Bottom: Autocorrelograms on the sustained firing rates of AIT cells along the serial position number of the stimuli. Abscissa is the relative position of the patterns in the training sequence, where patterns n,n+1 are first neighbors, and patterns n,n+2 are second neighbors. Triangles are mean correlations in responses to the learned stimuli for 57 cells. Open circles are correlations in responses to novel stimuli for 17 cells, and closed circles are responses to learned stimuli for the same 17 cells. Squares are mean correlations for the 28 cells with statistically significant response correlations, according to Kendall's correlation test. Adapted from Miyashita (1988). Reprinted with permission from MacMillan Magazines, Ltd., copyright 1988.

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). Földiák (Földiák, 1991) used temporal association to model the development of viewpoint invariant responses of complex V1 cells from sweeps of oriented edges across the retina. This model achieved translation invariance in a single layer by having orientation-tuned filters in the first layer that produced linearly separable patterns. More generally, approximate viewpoint invariance may be achieved by the superposition of several Földiák-like networks (Rolls, 1995). Most such models used idealized input representations. These learning mechanisms have recently been shown to learn transformation invariant responses to complex inputs such as images of faces (Bartlett and Sejnowski, 1996b; Bartlett and Sejnowski, 1997; Wallis and Rolls, 1997; Becker, 1999).

The assumption of temporal coherence can also be applied to learn other properties of the visual environment, such as depth from stereo disparity of curved surfaces (Becker, 1993; Stone, 1996).

There are several 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). Recurrent excitatory connections within a cortical area and reciprocal connections between cortical regions (O'Reilly and Johnson, 1994) could sustain activity over longer time periods and allow temporal associations across larger time scales.

The time course of the modifiable state of a neuron, based on the open time of the NMDA channel for calcium influx, has been modeled by a lowpass temporal filter on the post-synaptic unit activities (Rhodes, 1992). A lowpass temporal filter is a simple way to describe mathematically any of the above effects. This paper examines the contribution of such a lowpass temporal filter to the development of viewpoint invariant responses in both a feedforward layer, and a second, recurrent layer of a network. In the feedforward system, the Competitive Learning rule (Rumelhart and Zipser, 1985) is extended to incorporate an activity trace on the output unit activities (Földiák, 1991). The activity trace causes recently active output units to have a competitive advantage for learning subsequent input patterns.

The recurrent component of the simulation examines the development of temporal associations in an attractor network. Perceptual representations have been related to basins of attraction in activity patterns across an assembly of cells (Amit, 1995; Freeman, 1994; Hinton and Shallice, 1991). Weinshall and Edelman (Weinshall and Edelman, 1991) modeled the development of viewpoint invariant representations of wire-framed objects by associating neighboring views into basins of attraction. The simulations performed here show how viewpoint invariant representations of face images can be captured in an attractor network, and we examine the effect of a lowpass temporal filter on the attractor network learning rule. The recurrent layer was a generalization of a Hopfield network (Hopfield, 1982) with a lowpass temporal filter on all unit activities. We show that the combination of basic Hebbian learning with temporal smoothing of unit activities produces an attractor network learning rule that associates temporally proximal input patterns into basins of attraction. This learning rule is a generalization of an attractor network learning rule that produced temporal associations between randomly generated input patterns (Griniasty et al., 1993).

These two mechanisms were implemented in a model with both feedforward and lateral connections. The input to the model consisted of the outputs of an

array of Gabor filters. These were projected through feedforward connections to a second layer of units, where unit activities are passed through a lowpass temporal filter. The feedforward connections were modified by competitive Hebbian learning to cluster the inputs based on a combination of spatial similarity and temporal proximity. Lateral connections in the output layer created an attractor network that formed basins of attraction based on the temporal proximity of the input patterns. 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.

2. SIMULATION

Stimuli for these simulations consisted of 100 images of faces undergoing a change in pose, from David Beymer (Beymer, 1994) (see Figure 7.2). There were twenty individuals at each of five poses, ranging from $-30°$ to $30°$. The faces were automatically located in the frontal view image by using a feature-based template matching algorithm (Beymer, 1994). The location of the face in the frontal view image defined a window for the other images in the sequence. Each input sequence therefore consisted of a single stationary window within which the subject moved his or her head. The images were normalized for luminance and scaled to 120 x 120 pixels.

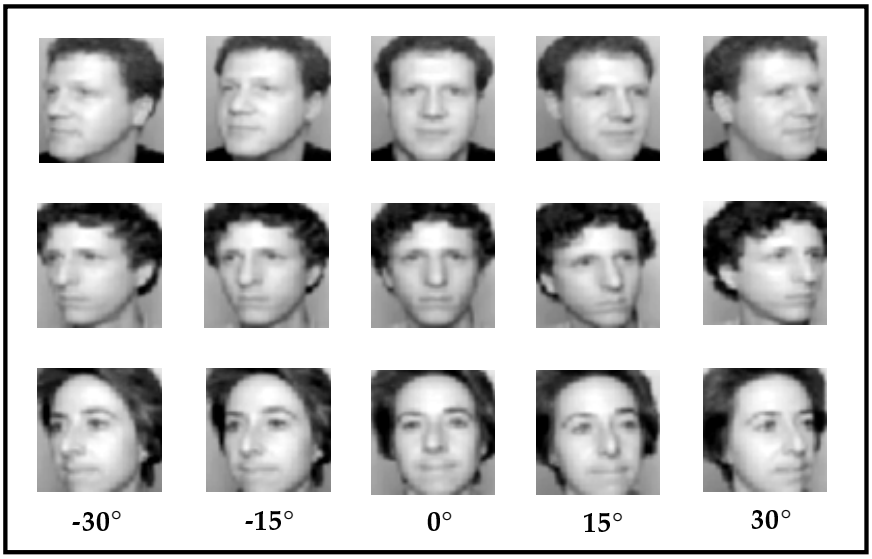

Figure 7.2. Sample of the 100 images used in the simulation. Image set provided by David Beymer (1994).

2.1. Model architecture

Images were presented to the model in sequential order as the subject changed pose from left to right (Figure 7.3). The first layer of processing consisted of an oriented energy model related to the output of V1 complex cells (Daugman, 1988; Lades et al., 1993). The images were filtered by a set of sine and cosine Gabor filters at 4 spatial scales (32, 16, 8, and 4 pixels per cycle), and at four orientations (vertical, horizontal, and $\pm 45^{\circ}$.) The standard deviation of the Gaussian was set to twice the frequency of the sine or cosine wave, such that the receptive field size of the spatial filters increased with the spatial scale of the filters. The outputs of the sine and cosine Gabor filters were squared and summed, and then normalized by scale and orientation (Heeger, 1991). The result was sampled at 8 pixel intervals. This produced a 3600-dimensional representation consisting of 225 spatial locations, 4 spatial scales, and 4 orientations.

The set of V1 model outputs projected to a second layer of 70 units labeled "complex pattern units" to characterize their receptive fields after learning. The complex pattern unit activities were passed through a lowpass temporal filter, described below. There was feedforward inhibition between the complex pattern units, meaning that the competition influenced the feedforward activations only. The 70 units were grouped into two inhibitory pools, such that there were two active complex pattern units for any given input pattern. The third stage of the model was an attractor network produced by lateral interconnections among all of the complex pattern units. The feedforward and lateral connections were updated successively.

2.2. Competitive Hebbian learning of temporal relationships

The learning rule for the feedforward connections of the model was an extension of the Competitive Learning Algorithm (Rumelhart and Zipser, 1985; Grossberg, 1976). The output unit activities were passed through a lowpass temporal filter (Bartlett and Sejnowski, 1996b). This manipulation gave active units in the previous time steps a competitive advantage for winning, and therefore learning, in the current time step.

Let $y_j^t = \sum_i w_{ij} x_i + b_j$ be the weighted sum of the feedforward inputs and the bias at time t. The activity of unit j at time t, $\overline{y_j}^{(t)}$, is determined by the trace, or running average, of its input activity:

$$\overline{y_j}^{(t)} = (1 - \lambda) y_j^t + \lambda \overline{y_j}^{(t-1)} \tag{7.1}$$

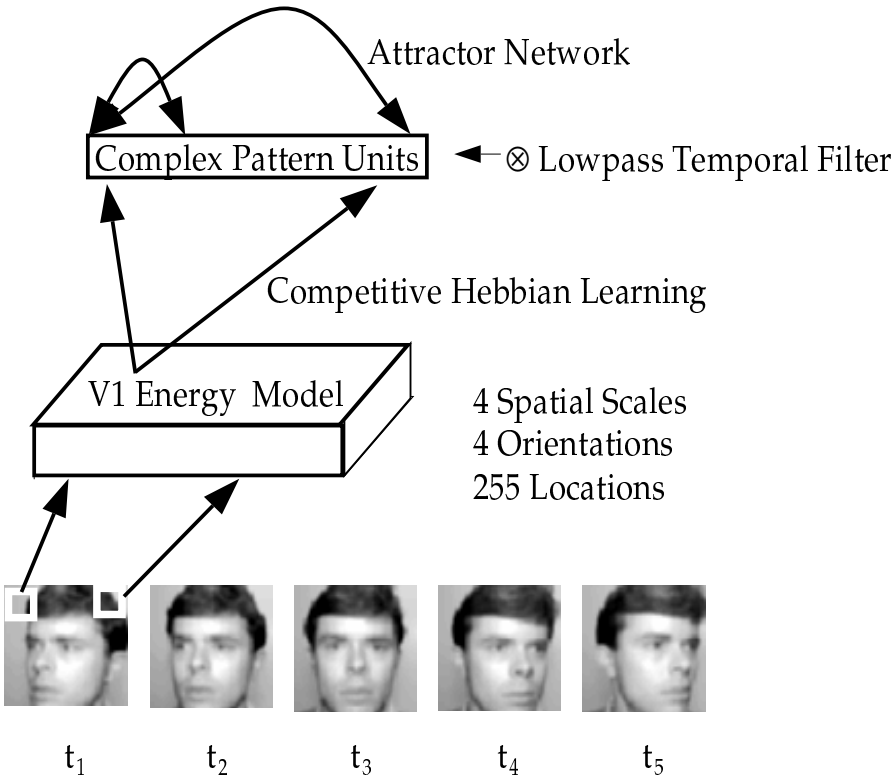

Figure 7.3. Model architecture.

The output unit activity, V_j, was subject to a step-nonlinear competition function.

$$V_j = \begin{cases} 1 & if j = max_j[\overline{y_j}^{(t)}] \\ \frac{\alpha}{N} & otherwise \end{cases} \tag{7.2}$$

where α is the learning rate, and N is the number of clustering units in the output layer. This was a modified winner-take-all competition where the non-winning activation was set to a constant small value rather than zero. The effect of the small positive activation was to cause non-winning weight vectors to move into the space spanned by the input data (Rumelhart and Zipser, 1985). The feedforward connections were updated according to the following learning rule:

$$\Delta w_{ij} = \alpha V_j (\frac{x_{iu}}{\sum_k x_{ku}} - w_{ij}) \tag{7.3}$$

The weight change from input i to output j was proportional to the normalized input activity at unit i for pattern u, x_{iu}, minus a weight decay term. In addition to the weight decay, the weight to each unit was constrained to sum to one by a divisive normalization.

The small positive activation of non-winning weight vectors does not guarantee that all weight vectors will eventually participate in the clustering. It causes the non-winning weight vectors to move slowly toward the centroid of the data, and some of the weight vectors may end up oscillating about the centroid without winning the competition for one of the inputs. A bias term was therefore added to cause each output unit to be active approximately the same proportion of the time. The learning rule for the bias to output unit j, b_j, was

$$\Delta b_j = \beta \left(\frac{P}{n} - c_j \right) \tag{7.4}$$

where P is the number of input patterns, n is the number of output units in one pool, and c_j is the count of wins for output j over the previous P time steps. The bias term was updated at the end of each iteration through the data, with learning rate β. If we define a unit's receptive field as the area of input space to which it responds, then the bias term acts to expand the receptive fields of units that tend to be inactive, and shrink the receptive fields of units that are active more often than the others. There is some justification for this due to evidence for activity dependent modification of receptive field size of cortical neurons, eg. (Jenkins et al., 1990; Kaas, 1991). An alternative way to normalize responses is through multiplicative scaling of the synaptic weights (Turrigiano et al., 1998).

One face image was input to the system per time step, so the face patterns, u, can also be indexed by the time step, t. The temporal smoothing was subject to reset based on discontinuities in optic flow, which insured that there was no temporal smoothing across input images with large changes. Optic flow between image pairs was calculated using a simple gradient-based flow estimator (Horn and Schunk, 1981). When the summed lengths of the optic flow vectors for sequential image pairs exceeded a threshold of $\gamma = 25$, $\overline{y}$ was initialized to y.[1] The competitive learning rule alone, without the temporal smoothing, partitioned the set of inputs into roughly equal groups by spatial similarity. With the temporal smoothing, this learning rule clustered the input

[1] This initialization is not strictly required for the success of such unsupervised learning algorithms because of the low probability of any specific pair of adjacent images of different individuals relative to the probability of adjacent images of the same individual, cf. (Wallis and Baddeley, 1997). However, we chose not to ignore the transitions between individuals since there are internal cues to these transitions such as eye movements, motion, and longer temporal delays.

by a combination of spatial similarity and temporal proximity, where the relative contribution of the two factors was determined by the parameter λ.

This learning rule is related to spatio-temporal principal component analysis. It has been shown that competitive Hebbian learning can find the first N principal components of the input data, where N is the number of output units (Oja, 1989; Sanger, 1989). The low-pass temporal filter on output unit activities in Equation 7.1 causes Hebbian learning to find axes along which the data covaries over recent *temporal* history. Due to the linear transfer function, passing the output activity through a temporal filter is equivalent to passing the input through the temporal filter. Competitive Hebbian learning can thus find the principal components of this spatio-temporal input signal.

2.3. Temporal association in an attractor network

The lateral interconnections in the output layer formed an attractor network. After the feedforward connections were established in the first layer using competitive learning, the weights of the lateral connections were trained with a basic Hebbian learning rule. Hebbian learning of lateral interconnections, in combination with the lowpass temporal filter (Equation 7.1) on the unit activities, produced a learning rule that associated temporally proximal inputs into basins of attraction. This is demonstrated as follows. We begin with a basic Hebbian learning algorithm:

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(y_i^t - y^0)(y_j^t - y^0) \tag{7.5}$$

where N is the number of units, P is the number of patterns, and y^0 is mean activity over all of the units. Replacing y_i^t with the activity trace $\overline{y_i}^{(t)}$ defined in Equation 7.1, we obtain

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}\left((1-\lambda)y_i^t + \lambda\overline{y_i}^{(t-1)} - y^0\right)\left((1-\lambda)y_j^t + \lambda\overline{y_j}^{(t-1)} - y^0\right) \tag{7.6}$$

Substituting $y^0 = \lambda y^0 + (1-\lambda)y^0$ and multiplying out the terms produces the following learning rule:

$$\begin{aligned} W_{ij} = \frac{1}{N}\sum_{t=1}^{P}\Big(& (1-\lambda)^2(y_i^t - y^0)(y_j^t - y^0) \\ & +\lambda(1-\lambda)\left[(y_i^t - y^0)(\overline{y_j}^{(t-1)} - y^0) + (\overline{y_i}^{(t-1)} - y^0)(y_j^t - y^0)\right] \\ & +\lambda^2\left[(\overline{y_i}^{(t-1)} - y^0)(\overline{y_j}^{(t-1)} - y^0)\right]\Big) \end{aligned} \tag{7.7}$$

This learning rule is a generalization of an attractor network learning rule that has been shown to produce correlated attractors based on serial position in the input sequence (Griniasty et al., 1993). The first term in this equation is basic Hebbian learning. The weights are proportional to the covariance matrix of the input patterns at time t. The second term performs Hebbian association between the patterns at time t and $t-1$. The third term is Hebbian association of the trace activity for pattern $t-1$.

The following update rule was used for the activation V of unit i at time t from the lateral inputs (Griniasty et al., 1993):

$$V_i(t+\delta t) = \phi\left[\sum W_{ij}V_j(t) - \theta\right] \tag{7.8}$$

Where θ is a neural threshold and $\phi(x) = 1$ for $x > 0$, and 0 otherwise. In these simulations, $\theta = 0.007$, $N = 70, P = 100$, $y^0 = 0.03$, and $\lambda = 0.5$.

The learning rule developed by Griniasty, Tsodyks, and Amit (Griniasty et al., 1993) is presented in Equation 7.9 for comparison. The Griniasty et. al. learning rule associates first neighbors in the pattern sequence, whereas the learning rule in 7.7 has a longer memory. The weights in 7.9 are a function of the *discrete* activities at t and $t-1$, whereas the weights in 7.7 are a function of the current input and the activity *history* at time $t-1$.

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(y_i^t - y^0)(y_j^t - y^0) +$$

$$a\left[(y_i^{t+1} - y^0)(y_j^t - y^0) + (y_i^t - y^0)(y_j^{t+1} - y^0)\right] \tag{7.9}$$

The weight structure and fixed points of an attractor network trained with Equation 7.7 are illustrated in Figures 7.4 and 7.5 using an idealized data set in order to facilitate visualization. The fixed points for the real face data will be illustrated later, in Section 2.4. The idealized data set contained 25 input patterns, where each pattern was coded by activity in a single bit (Figure 7.4, Top). The patterns represented 5 individuals with 5 views each (a - e). The middle graph in Figure 7.4 shows the weight matrix obtained with the attractor network learning rule, with $\lambda = 0.5$. Note the approximately square structure of the weights along the diagonal, showing positive weights among most of the 5 views of each individual. The inset shows the actual weights between views of individuals 3 and 4. The weights decrease with the distance between the patterns in the input sequence. The bottom graphs show the sustained patterns of activity in the attractor network for each input pattern. Unlike the standard Hopfield network, in which the objective is to obtain sustained activity patterns that are identical to the input patterns, the objective here is to have a many-to-one mapping from the five views of an individual to a single

pattern of sustained activity. Note that the same pattern of activity is obtained no matter which of the 5 views of the individual is input to the network. For this simplified representation, the attractor network produces responses that are entirely viewpoint invariant. The fixed points in this demonstration are the conjunctions of the input activities for each individual view.

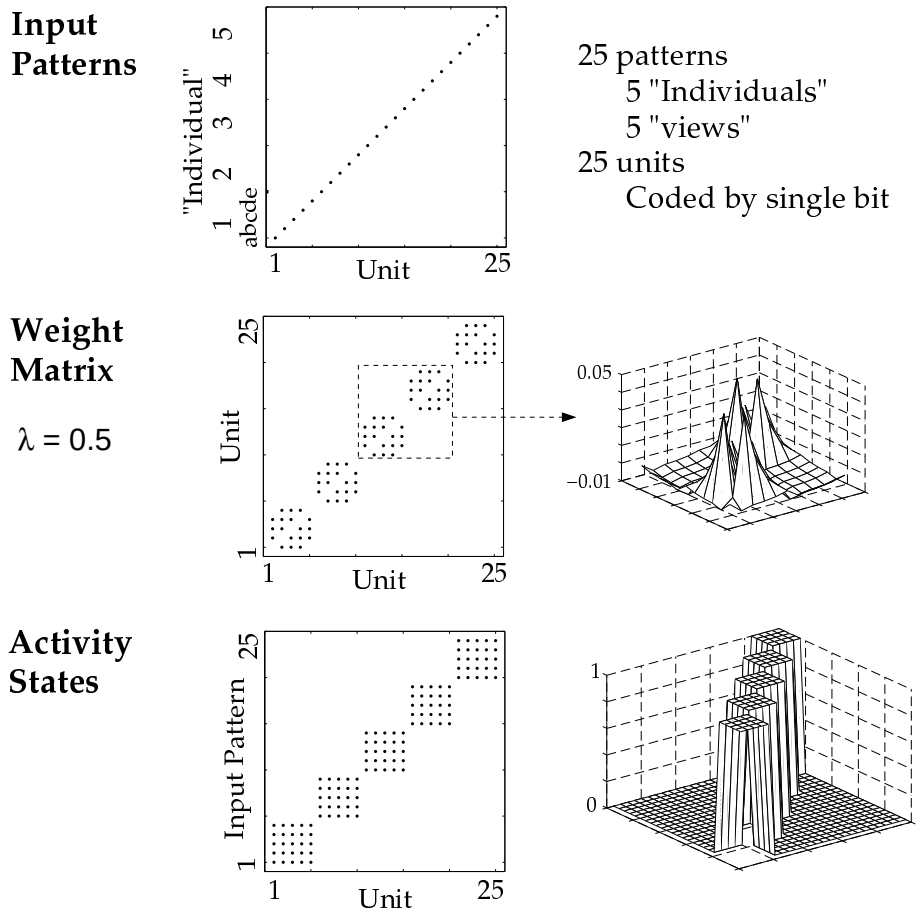

Figure 7.4. Demonstration of attractor network with idealized data. Top: Idealized data set. The patterns consist of 5 "individuals" (1,2,3,4,5) with five "views" each (a,b,c,d,e), and are each coded by activity in 1 of the 25 units. Center: The weight matrix obtained with equation 3. Dots show the locations of positive weights, and the inset shows the actual weights among the 5 views of two different individuals. Bottom: Fixed points for each input pattern. Unit activities are plotted for each of the 25 input patterns.

Figure 7.5 shows the weight matrix for different values of the temporal filter, λ.[2] As λ increases, a larger range of views contain positive weights. The figure also gives the fixed points for each input pattern. For $\lambda = 0.25$, 2 to 3 views are associated into the same basin of attraction. For $\lambda = 0.4$, there are positive connections between only a subset of the views for each face, yet this weight matrix is sufficient to associate all five views into the same basin of attraction. A rigorous numerical analysis of the mean field equations and fixed points of a related weight matrix can be found in (Parga and Rolls, 1998).

2.4. Simulation results

Sequences of graylevel face images were presented to the network in order as each subject changed pose. Faces rotated from left to right and right to left in alternate sweeps. The feedforward and the lateral connections were trained successively. The feedforward connections were updated by the learning rule in Equations 7.1-7.3, with $\lambda = 0.5$. Competitive interactions were among two pools of 35 units so that there were two active outputs for each input pattern. The two competitive pools created two samples of image clustering, which provided additional information on relationships between images. Images could be associated by both clusters, one, or neither, and images that were never clustered together could share a common clustering partner.

After training the feedforward connections, the representation of each face was a sparse representation consisting of the two active output units out of the total of 70 complex pattern units. "Pose tuning" of the feedforward system was assessed by comparing correlations in the network outputs for different views of the same face to correlations across faces of different people. Mean correlations for different views of the same face were obtained for each possible change in pose by calculating mean correlation in feedforward outputs across all four 15^0 changes in pose, three 30^0 changes in pose, and so forth. Mean correlations *across* faces for the same changes in pose were obtained by calculating mean correlation in feedforward outputs for different subjects across all 15^0 changes in pose, 30^0 changes in pose, and so forth.

Figure 7.6 (Top Left) shows pose tuning both with and without the temporal lowpass filter on unit activities during training. The temporal filter broadened the pose tuning of the feedforward system, producing a response that was more selective for the individual and less dependent on viewpoint.

The discriminability of the feedforward output for same-face versus different-face was measured by calculating the receiver-operator-characteristic (ROC) curve for the distributions of same-face and different-face output correlations. An ROC curve plots the proportion of hits against the proportion of false alarms

[2]The half-life, h, of the temporal filter is related to λ by $\lambda^h = 0.5$ (Stone, 1996). For $\lambda = 0.5$, the activity at time t is reduced by 50% after one time step.

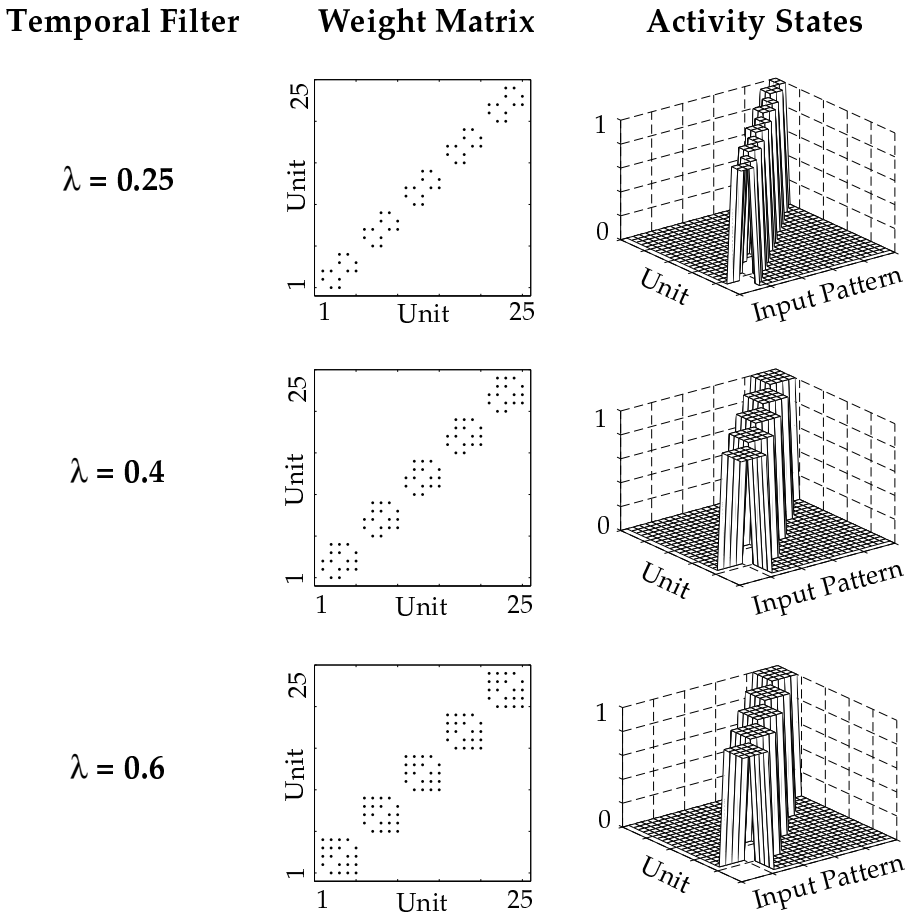

Figure 7.5. Weight matrix (left) and fixed points (right) for three values of the temporal filter, λ. Dots show locations of positive weights. Unit activities are plotted for each of the 25 input patterns of the simplified data.

(FA's) for choosing between two distributions at different choices of the acceptance criteria. The area under the ROC measures the discriminability of the two distributions, ranging from 0.5 for fully overlapping distributions to 1.0 for distributions with zero overlap in the tails. Figure 7.6 (Top Right) shows the ROC curves and areas under the ROC for feedforward output correlations with $\lambda = 0.5$ and $\lambda = 0.0$. The temporal filter increased the discriminability of the feedforward outputs.

Test image results were obtained by alternately training on four poses and testing on the fifth, and then averaging across all test cases. Test images

Table 7.1. Contribution of the feedforward connections and the attractor network to viewpoint invariance of the complete system. Area under the ROC for the sustained activity patterns in network layer 2 is given with and without the temporal activity trace in during learning in the feedforward connections (λ_1) and in the attractor network (λ_2).

	λ_1	
λ_2	0	0.5
0	.70	.90
0.5	.84	.98

produced a similar pattern of results, which are presented in the bottom of Figure 7.6.

The feedforward system provided a sparse input to the attractor network. After the feedforward connections were established, the feedforward weights were held fixed, and sequences of face images were again presented to the network as each subject gradually changed pose. The lateral connections among the output units were updated by the learning rule in Equation 7.7. After training the attractor network, each face was presented to the system, and the activities in the output layer were updated until they arrived at a stable state. The sustained patterns of activity comprised the representation of a face in the attractor network component of the model. Following learning, these patterns of sustained activity were approximately viewpoint invariant.

Figure 7.7 shows pose tuning and ROC curves for the sustained patterns of activity in the attractor network. The graphs compare activity correlations obtained using five values of λ in Equation 7.7. Note that $\lambda = 0$ corresponds to a standard Hebbian learning rule. The contribution of the feedforward system and the attractor network to the overall viewpoint invariance of the system are compared in Table 7.1. Temporal associations in the feedforward connections and the lateral connections both contributed to the viewpoint invariance of the sustained activity patterns of the system.

Figure 7.8 shows the activity in network layer two for 25 of the 100 graylevel face images, consisting of five poses of five individuals. Face representations following training of the feedforward connections only with $\lambda = 0$ (top) are contrasted with face representations obtained when the feedforward connections were trained with $\lambda = 0.5$ (middle), and with the face representations in the attractor network, in which both the feedforward and lateral connections were trained with $\lambda = 0.5$. Competitive Hebbian learning without the temporal lowpass filter frequently included neighboring poses of an individual in a cluster, but the number of views of an individual within the same cluster did not exceed two, and the clusters included images of other individuals as well. The temporal lowpass filter increased the number of views of an individual within a

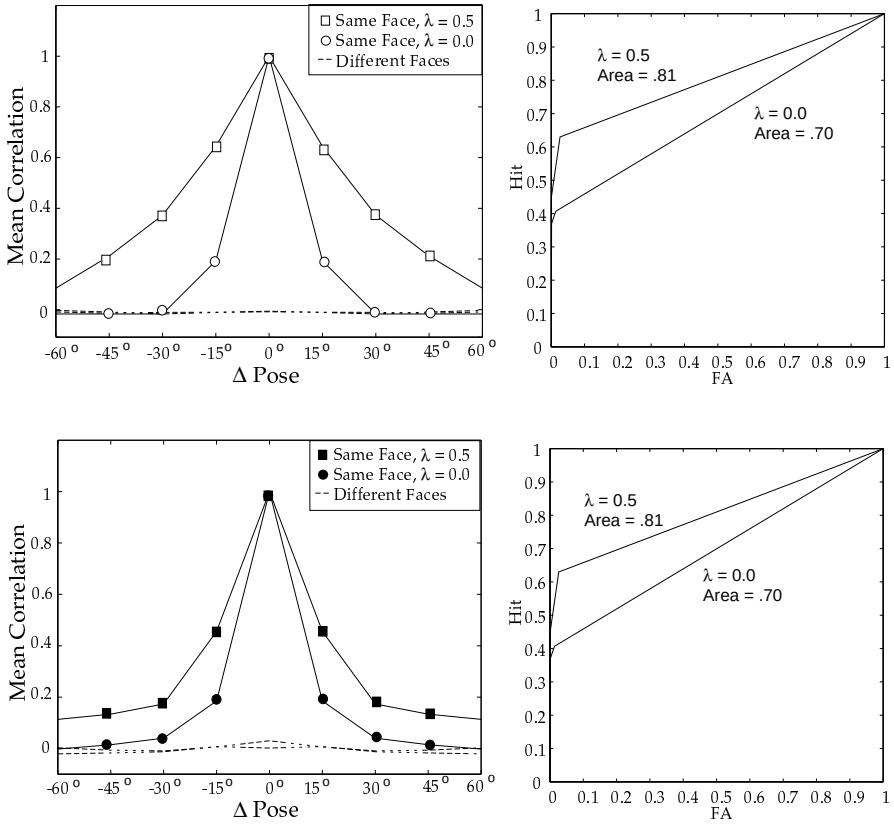

Figure 7.6. Pose tuning and ROC curves of the feedforward system for training images (top) and test images (bottom). Left: Mean correlations of the feedforward system outputs for pairs of face images are presented by change in pose. Correlations across different views of the same face (–) are compared to correlations across different faces (−−) for two values of the temporal trace parameter $\lambda = 0.5$ and $\lambda = 0$. Right: ROC curves and area under the ROC for same face vs. different face discrimination of the feedforward system outputs for training images (top) and test images (bottom).

cluster. Note however, that for individuals 4 and 5, the representation of views *a* and *b* are not correlated with that of views *d* and *e*. The attractor network of the bottom plot was trained on the face codes shown in the middle plot, with $\lambda = 0.5$. The attractor network increased the correlation in face codes for different views of an individual. In the sample shown, the representations for individuals 1 - 4 became viewpoint invariant, and the representations for the views of individual 5 became highly correlated. Consistent with the findings of Weinshall & Edelman (Weinshall and Edelman, 1991) for idealized wire-framed objects, units that were active for one view of a face in the input to the

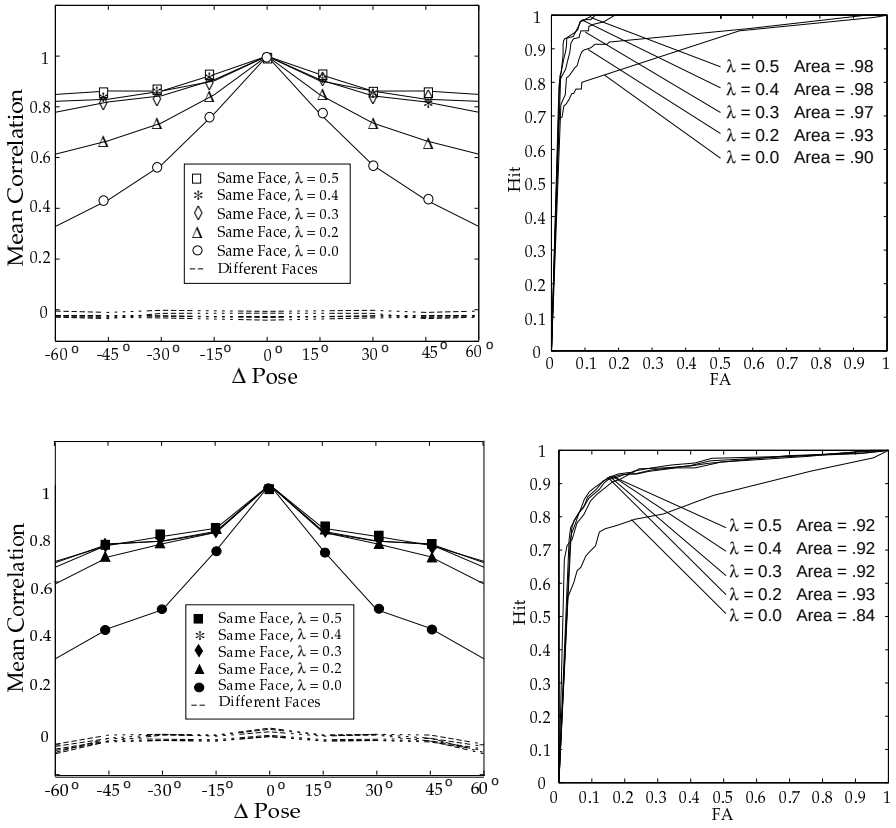

Figure 7.7. Pose tuning and ROC curves of the attractor network for training images (top) and test images (bottom). Left: Mean correlations in sustained activity patterns in the attractor network for pairs of face images are presented by change in pose. Correlations across different views of the same face (–) are compared to correlations across different faces (−−) for five values of the temporal trace parameter λ. Right: ROC curves and area under the ROC for same face vs different face discrimination of the sustained activity patterns for training images (top) and test images (bottom).

attractor network exhibited sustained activity for more views, or all views of that face in the attractor network.

The storage capacity of this attractor network, defined as the maximum number of individual faces that can be stored and retrieved in a view-invariant way, F_{max}, depends on several factors. These include the load parameter, $\frac{P}{N}$, where P is the number of input patterns and N is the number of units, the number of views, s, per individual, and the coding efficiency, or sparseness, y_0. A detailed analysis of the influence of these factors on capacity has been

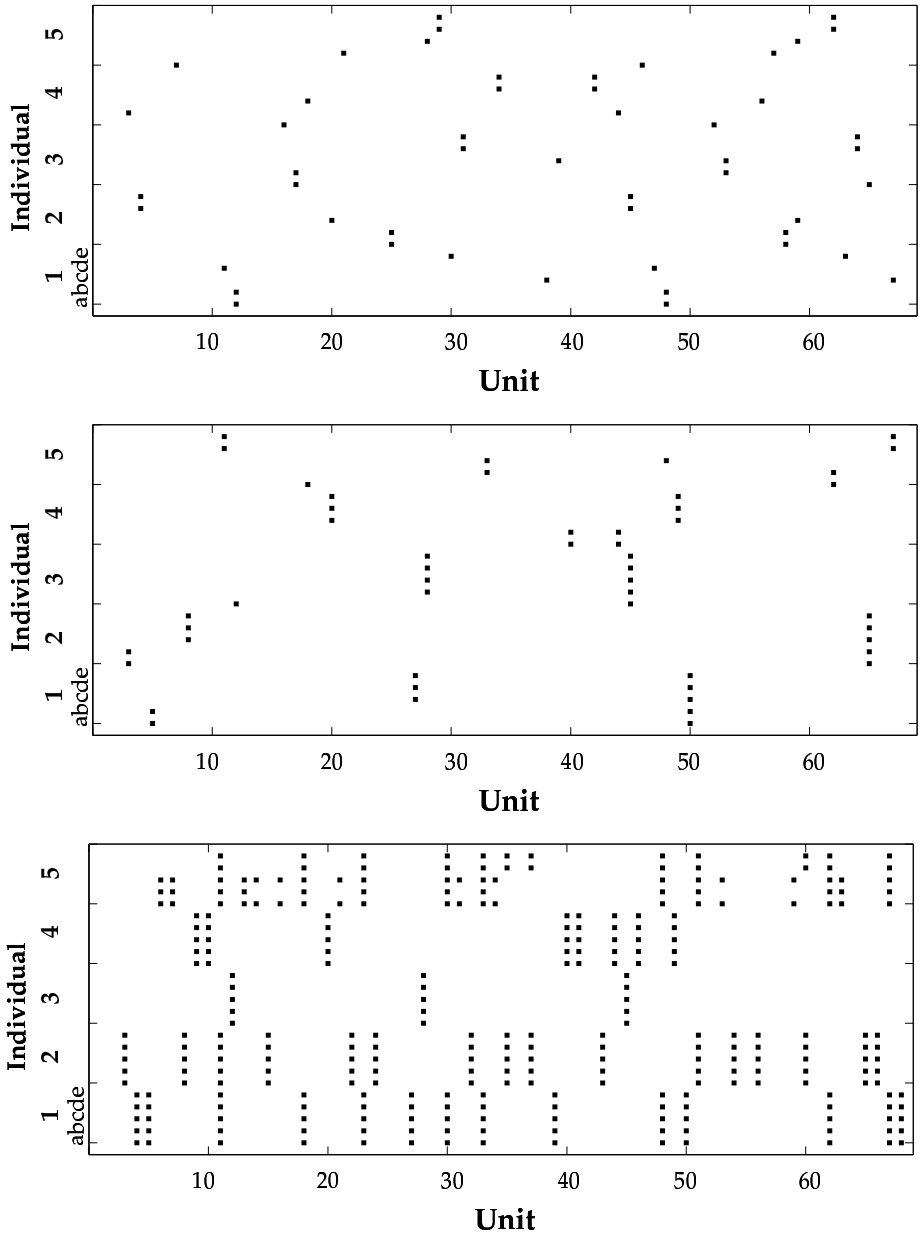

Figure 7.8. Coding of real face image data. Top: Coding of 5 faces in network layer 2 following training of the feedforward connections only, with no temporal lowpass filter ($\lambda = 0.$) The vertical axis is the input image, with the five poses of each individual labeled a,b,c,d,e. The two active units for each input image are indicated on the horizontal axis. Middle: Coding of the same five faces following training of the feedforward connections with $\lambda = 0.5$. Bottom: Sustained patterns of activity in the attractor network for the same five faces, where both the feedforward and the lateral connections were trained with $\lambda = 0.5$.

presented elsewhere (Parga and Rolls, 1998). See also (Gardner, 1988; Tsodyks and Feigel'man, 1988).

We will outline some of these influences here. It has been shown for the autoassociative Hopfield network, for which the number of fixed points equals the number of input patterns, that the network becomes unstable with $\frac{P}{N} > 0.14$ (Hopfield, 1982). For the present network, we desired one fixed point per individual, where there were $s = 5$ input patterns per individual. Thus the capacity depended on $\frac{F}{N}$, where $F = \frac{P}{s}$ was the number of individuals in the input. The capacity of the attractor network also depended on the sparseness, y_0, since capacity increases as the mean activity level decreases according to $(y_0|ln(y_0)|)^{-1}$ (Gardner, 1988; Tsodyks and Feigel'man, 1988). Specifically, the capacity of attractor networks with $\{0, 1\}$ coding and s input patterns per desired memory depends on the number of neurons, N, and the sparseness of the input patterns, y_0, in the following way (Tsodyks and Feigel'man, 1988; Parga and Rolls, 1998):

$$\frac{F}{N} \leq \frac{0.2}{s^2 y_0 \ ln\left(\frac{1}{sy_0}\right)} \tag{7.10}$$

For the network with $N = 70$ units, sparseness $y_0 = 0.029$, and $s = 5$ views per individual, the maximum load ratio was $\frac{F}{N} = 0.14$, and the maximum number of individuals that can be stored in separate basin of attraction was $F_{max} = 10$.

Since storage capacity in the attractor network depends on coding efficiency, the proportion of active input units per pattern, the attractor network component of the model required its input representations to be sparse. Sparse inputs may be an appropriate assumption, given the sparseness of responses reported in V4 (Gallant et al., 1994) and area TE, a posterior IT region which projects to the anterior IT regions where transformation invariant responses can be found (Tanaka, 1993). The representations of faces in the attractor network itself were less sparse than its input, with a mean unit activity of 0.19 for each face, compared to 0.03 for its input, and each unit participated in the coding of 13 of the 100 images on average in the attractor network, compared to 3 images for its input. The coding levels in the attractor network were consistent with the sparse-distributed face coding reported in IT (Young and Yamane, 1992; Abbott et al., 1996).

We evaluated face recognition performance of the attractor network using a nearest neighbor classifier on the sustained activity patterns at several loading levels. Table 7.2 gives percent correct recognition performance of the sustained activity patterns in the network trained on real face data. Test patterns were assigned the class of the pattern that was closest in Euclidean distance. Each pattern was taken in turn as a test pattern and compared to the other 99, and then a mean was taken across the 100 test cases. Classification performance depended

		Attractor Network			***Eigenfaces***
F	P	N	$\frac{F}{N}$	%Correct	% Correct
5	25	70	.07	100	100
10	50	70	.14	90	90
20	100	70	.29	61	87

Table 7.2. Nearest neighbor classification performance of the attractor network. F: Number of individuals; P: Number of input patterns; N: Number of units. Classification performance is presented for three values of the load parameter, $\frac{F}{N}$. Results are compared to Eigenfaces for the same subset of faces. Classification performance of the attractor network is good when $\frac{F}{N} < 0.14$

on the load parameter, $\frac{F}{N}$. Performance was quite good when $\frac{F}{N} << 0.14$, and decreased as $\frac{F}{N}$ increased beyond this value. Classification errors occurred when two or more individuals shared a single basin of attraction.

Classification performance of the network for $F = 10$ was below 100% because not all fixed points were found. The set of input patterns did not cover all 10 basins of attraction. Since the input patterns (the outputs of the feedforward system) were driven by real face images, the input patterns were not constrained to be orthogonal. When the input patterns were orthogonal, such as the idealized data in Figure 7.4 in which each input was coded by activity in a different unit, then all fixed points were found for $F = F_{max}$ individuals, and classification performance was 100%.

3. DISCUSSION

Many cells in the primate anterior inferior temporal lobe and superior temporal sulcus maintain their response preferences to faces or three-dimensional objects over substantial changes in viewpoint (Hasselmo et al., 1989; Perrett et al., 1989; Logothetis and Pauls, 1995). This set of simulations demonstrated how such viewpoint invariant representations of faces could be developed from visual experience through unsupervised learning.

The inputs to the model were similar to the responses of V1 complex cells, and the goal was to apply unsupervised learning mechanisms to transform these inputs into pose invariant responses. We showed that a lowpass temporal filter on unit activities, which has been related to the time course of the modifiable state of a neuron (Rhodes, 1992), cooperates with Hebbian learning to (1) increase the viewpoint invariance of responses to faces in a feedforward system, and (2) create basins of attraction in an attractor network which associate temporally proximal inputs. This simulation demonstrated how viewpoint invariant representations of complex objects such as faces can be developed from

visual experience by accessing the temporal structure of the input. The model addressed potential roles for both feedforward and lateral interactions in the self-organization of object representations, and demonstrated how viewpoint invariant responses can be learned in an attractor network.

Temporal sequences contain information that can aid in the process of representing and recognizing faces and objects. 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) found that 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. The dynamic signal therefore contributed to the object representation beyond providing structure-from-motion. This model in this paper presented a means by which temporal information can be incorporated in the representation of a face.

Related models that have been developed independently support the results presented in this paper. Wallis and Rolls (Wallis and Rolls, 1997) trained a hierarchical feedforward system using Hebbian learning and the temporal activity trace of Equation 7.1. Their system successfully learned translation invariant representations of seven faces, and rotation invariant representations of three faces. Parga and Rolls (Parga and Rolls, 1998) presented a detailed analysis of the phase transitions and capacity of an attractor network related to the recurrent layer of the present network. Their work focused on the thermodynamic properties of this attractor network, using a predefined coupling matrix and idealized stimuli. Our work extends this analysis to the learning mechanisms that could give rise to such a weight matrix, and implements them in a system taking real images of faces as input.

The feedforward processing in this model was related to spatio-temporal principal component analysis of the Gabor filter representation. It has been shown that competitive Hebbian learning finds the principal components of the input data (Oja, 1989; Sanger, 1989). The learning rule in the feedforward component of this model extracted information about how the Gabor filter outputs covaried in recent temporal history in addition to how they covaried over static views.

In this model, pose invariant face recognition was acquired by learning associations between 2-dimensional patterns, without recovering 3-D coordinates or structural descriptions. It has been proposed that 3-D object recognition may not require explicit internal 3-dimensional models, as was previously assumed, and recognition of novel views may instead be accomplished by linear (Ullman and Basri, 1991) or nonlinear combination 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).

Further support for view-based representations comes from a related model that simulated "mental rotation" response curves in a system that stored multiple 2-dimensional views and their temporal associations (Weinshall and Edelman, 1991). Weinshall and Edelman trained a 2 layer network to store individual views of wire-framed objects, and then updated lateral connections in the output layer with Hebbian learning as the input object rotated through different views. The strength of the association 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, one view of an object was presented and the output activity was iterated until all of the units for that object were active. 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, mimicking the linear increase in human response times that has been taken has evidence for mental rotation of an internal 3-D model (Shepard and Cooper, 1982).

In example-based models of recognition such as radial basis functions (Poggio and Edelman, 1990), neurons with view-independent responses are proposed to pool responses from view-dependent neurons. Our model suggests a mechanisms for how this pooling could be learned. Logothetis and Pauls (Logothetis and Pauls, 1995) reported a small percentage of viewpoint invariant responses in the AIT of monkeys that were trained to recognize wire-framed objects across changes in view. The training images in this study oscillated $\pm 10^0$ from the vertical axis. The temporal association hypothesis presented in this paper suggests that more viewpoint invariant responses would be recorded if the monkeys were exposed to full rotations of the objects during training.

Acknowledgments

This project was supported by Lawrence Livermore National Laboratory ISCR Agreement B291528, and by the McDonnell-Pew Center for Cognitive Neuroscience at San Diego. We thank Tomaso Poggio, James Stone, and Laurenz Wiskott for valuable discussions on earlier drafts of this paper.

Chapter 8

CONCLUSIONS AND FUTURE DIRECTIONS

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 driven by the dependencies (Barlow, 1989). According to Barlow's theory, what is important for a system to be able to detect is new regularities that differ from the environment to which the system has been adapted. These are what Barlow refers to as "suspicious coincidences." Learning mechanisms that encode the dependencies that are expected in the input and remove them from the output better enable a system to detect these new regularities in the environment. Independence facilitates the detection of high-order relationships that characterize an object because the prior probability of any particular high order combination of features 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. A number of unsupervised learning algorithms have been devised that attempt to learn the structure of the input by employing an objective of reducing statistical dependencies between coding elements.

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) "holons" (Cottrell and Metcalfe, 1991), and "local feature analysis" (Penev and Atick, 1996) are data-driven face representations based on principal component analysis. 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. Principal component analysis separates the correlations in the input, but does not address the high order dependencies such as the relationships among

three or more pixels. In a task such as face recognition, much of the important information may be contained in these high-order dependencies.

Independent component analysis 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 was recently developed (Bell and Sejnowski, 1995). This algorithm is an unsupervised learning rule derived from the principle of optimal information transfer through sigmoidal neurons (Laughlin, 1981; Atick, 1992), and information maximization (Linsker, 1988). The algorithm maximizes the mutual information between the input and the output of a transfer function, which produces statistically independent outputs under certain conditions. Independent component analysis does not constrain the axes to be orthogonal, and attempts to place them in the directions of statistical dependencies in the data. Each weight vector in ICA attempts to encode a portion of the input dependencies in order to remove the redundancies from *between* the inputs and transform them into redundancies *within* the response distributions of the individual output units.

Chapter 3 developed representations for face recognition based on statistically independent components of face images. The information maximization 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 an independent feature code across images. Face recognition performances with the ICA representations were compared to the Eigenface approach, which is based on PCA. Both ICA representations were superior to the PCA representation for recognizing faces across sessions and changes in expression. A combined classifier that took account of the image similarities within both ICA representations outperformed PCA for all conditions tested. We have demonstrated elsewhere that ICA representations can outperform PCA representations for recognizing faces across changes in pose, and changes in lighting (Bartlett and Sejnowski, 1997).

Chapters 5 and 6 compared image representations for facial expression analysis, and demonstrated that representations derived from redundancy reduction on the graylevel face image ensemble are powerful for face image analysis. The independent component representation described above was compared to a number of other face image representation algorithms for recognizing facial actions in a project to automate the Facial Action Coding System (Ekman and Friesen, 1978). Chapter 5 showed that a PCA representation 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 to more than eight other image representations, including 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 principal component analysis. 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 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 (Simoncelli, 1997) that Gabor filter outputs of natural images are pairwise independent in the presence of divisive normalization similar to the length normalization in the Gabor representation of Chapter 6.

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 examined unsupervised learning of viewpoint invariant representations of faces through spatio-temporal redundancy reduction. This work explored the development of viewpoint invariant responses to faces from visual experience in a biological system. Through coding principles that are sensitive to temporal redundancy in the input in addition to spatial redundancy, it is possible to learn viewpoint invariant representations. In natural visual experience, 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. A set of simulations demonstrated how viewpoint invariant representations of faces can be developed from visual experience by capturing the temporal relationships among the input patterns. The simulations explored the interaction of temporal smoothing of activity signals with Hebbian learning (Földiák, 1991) in both a feed-forward system and a recurrent system. The recurrent system was a generalization of a Hopfield network with a lowpass temporal filter on all unit activities. 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 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 gave better performance for recognizing facial identity than a representation that separated only the second-order dependencies. In addition, learning strategies that encoded the spatio-temporal redundancies in the input extracted structure relevant to visual invariances.

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). The research in this book demonstrated that such mechanisms are also effective for high level visual recognition tasks such as facial identity recognition and expression recognition.

As discussed in Chapters 2 and 3, the ICA algorithm produces sparse outputs (Bell and Sejnowski, 1997). Due to the advantages of sparse codes for associative memory (Baum et al., 1988), the ICA factorial representation would constitute a good input representation for the attractor network model of Chapter 7. Preliminary explorations demonstrated the success of this implementation for learning pose invariant representations of faces (Bartlett and Sejnowski, 1996a).

The spatio-temporal model of Chapter 7 focused on learning second-order redundancies via Hebbian learning. Future directions for this research include exploring spatio-temporal independent component analysis for learning visual invariances. One method for extracting spatio-temporal independent components is to perform ICA on image sequences, where the concatenated video frames of a face changing pose are treated as a single sample (Sone et al., 1999). Alternatively, methods for extracting the spatio-temporal independent components of a dataset X in which one dimension is space and the other dimension is time are currently under development (Sone et al., 1999). In a similar vein, a recent review (Peruš, 2001) proposed phase-Hebbian learning as a means for incorporting high-order information in the attractor network learning rule of Chapter 7. Another way to incorporate high-order information in the attractor network learning rule is by implementing a quantum associative network (Peruš, 2000), in which the real valued variables in a Hopfield network are translated into quantum complex valued (phase information carrying) values.

Another area for exploration is methods for extracting fewer sources than mixtures for the independent component representations presented in Chapter 3. In Chapter 3, the number of sources was controlled by reducing the dimensionality of the data through principal component analysis prior to per-

forming ICA. There are two limitations to this approach (Stone and Porrill, 1998). The first is the reverse dimensionality problem. It may not be possible to linearly separate the independent sources in smaller subspaces. Since the analysis in Chapter 3 retained a reasonably high dimensionality (200), this may not have been a serious limitation of this approach. Secondly, it may not be desirable to throw away subspaces of the data with low power such as the higher principal components. Although low in power, these subspaces may contain independent components, and the property of the data we seek is independence, not amplitude. Techniques have been proposed for separating sources on projection planes without discarding any independent components of the data (Stone and Porrill, 1998).

The information maximization algorithm employed to perform independent component analysis in this thesis assumed that the underlying "causes" of the pixel graylevels in face images had a super-Gaussian (peaky) response distribution. Many natural signals, such as sound sources, have been shown to have a super-Gaussian distribution (Bell and Sejnowski, 1995). The underlying "causes" of the pixel graylevels in the face images are unknown, and it is possible that some of the causes could have had a sub-Gaussian distribution. Any sub-Gaussian sources would have remained mixed. Methods for separating sub-Gaussian sources through information maximization have recently been developed (Lee et al., 1999), and another future direction is to examine sub-Gaussian components of face images.

The information maximization algorithm employed in this thesis also assumed that the pixel values in face images were generated from a mixing process that could be linearly approximated. This linear approximation has been shown to hold true for the effect of lighting on face images (Hallinan, 1995). Other influences, such as changes in pose and expression, have nonlinear effects. Although the effects of *small* changes in pose and expression may be linearly approximated, an algorithm for extracting nonlinear independent components may be better suited to representing these contributions to the pixel values. Nonlinear independent component analysis in the absence of prior constraints is an ill-conditioned problem, but some progress has been made by assuming a linear mixing process followed by parametric nonlinear functions (Lee et al., 1997).

A second approach to independent component analysis involves building a generative model of the data using maximum likelihood methods (MacKay, 1996). Each data point x is assumed to be a linear mixture of independent sources, $x = As$, where A is a mixing matrix, and s contains the sources. A likelihood function of the data can then be generated under this model, with the assumption that the sources s are independent. The elements of the basis matrix A and the sources s can then be obtained by gradient ascent on the log likelihood function. Factors that combine nonlinearly to influence the pixel

graylevels such as pose and lighting can be separated with in this framework as follows (David Mackay, personal communication). Each source, s_i can be modeled as a nonlinear combination of other sources. For example s could me modeled as a multiplicative interaction of a pose parameter p and a lighting parameter l by $s_i = p_i l_i$. The maximum likelihood problem then becomes one of maximizing $P(x|p, l, A)$, where the products $s_i = p_i l_i$ are assumed to be independent.

An alternative method for representing the face images that can accommodate nonlinear mixtures of sources is to learn an "overcomplete" basis set (Lewicki and Olshausen, 1998). In this representation, more bases are learned than are necessary to completely describe the data, hence the term "overcomplete." Overcomplete bases can be learned from a generalization of the maximum likelihood ICA algorithm, and can result in codes that are a nonlinear function of the data. Although a complete basis is sufficient to describe the data, overcomplete bases are better able to capture the underlying structure of complicated data distributions.

References

Abbott, L., E., R., and Tovee, M. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6(3):498–505.

Adolphs, R., Damasio, H., Tranel, D., and Damasio, A. (1996). Cortical systems for the recognition of emotion in facial expressions. *Journal of Neuroscience*, 16(23):7678–7687.

Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. (1995). Fear and the human amygdala. *Journal of Neuroscience*, 15(9):5879–5891.

Amari, s., Cichocki, A., , and Yang, H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, Cambridge, MA. MIT Press.

Amit, D. (1995). The hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences*, 18:617–657.

Antonini, A. and Stryker, M. (1993). Development of individual geniculocortical arbors in cat striate cortex and effects of binocular impulse blockade. *Journal of Neuroscience*, 13(8):3549–73.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251.

Atick, J. and Redlich, A. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Ballard, D. (1997). *An introduction to natural computation.* MIT Press, Cambridge, MA.

Barlow, H. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Barlow, H. (1994). What is the computational goal of the neocortex? In Koch, C., editor, *Large scale neuronal theories of the brain*, pages 1–22. MIT Press, Cambridge, MA.

Bartlett, M. (1998). *Face Image Analysis by Unsupervised Learning and Redundancy Reduction.* PhD thesis, University of California, San Diego.

Bartlett, M., Donato, G., Movellan, J., Hager, J., Ekman, P., and Sejnowski, T. (2000). Image representations for facial expression coding. In Solla, S., Leen,

T., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–263.

Bartlett, M., Lades, H., and Sejnowski, T. (1998). Independent component representations for face recognition. In Rogowitz, B. & Pappas, T., editor, *Proceedings of the SPIE Symposium on Electonic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, San Jose, CA. SPIE Press.

Bartlett, M. and Sejnowski, T. (1996a). learning viewpoint invariant representations of faces in an attractor network. In Cottrell, G., editor, *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, page 730.

Bartlett, M. and Sejnowski, T. (1996b). Unsupervised learning of invariant representations of faces through temporal association. In Bower, J., editor, *Computational Neuroscience: Trends in Research; Int. Rev. Neurobio. Suppl. 1*, pages 317–322, San Diego, CA. Academic Press.

Bartlett, M. and Sejnowski, T. (1997). Viewpoint invariant face recognition using independent component analysis and attractor networks. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 817–823, Cambridge, MA. MIT Press.

Bartlett, M., Viola, P., Sejnowski, T., Larsen, J., Hager, J., and Ekman, P. (1996). Classifying facial action. In Touretski, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 823–829. Morgan Kaufmann, San Mateo, CA.

Bassili, J. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059.

Baum, E., Moody, J., and Wilczek, F. (1988). Internal representaions for associative memory. *Biological Cybernetics*, 59:217–228.

Becker, S. (1992). *An Information-theoretic Unsupervised Learning Algorithm for Neural Networks*. PhD thesis, University of Toronto.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In Hanson, S., Cowan, J., and Giles, C., editors, *Advances in Neural Information Processing Systems*, volume 5, pages 361–368, San Mateo, CA. Morgan Kaufmann.

Becker, S. (1999). Implicit learning in 3d object recognition: The importance of temporal context. *Neural Computation*, 11(2):347–74.

Becker, S. and Hinton, G. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 335(6356):161–3.

Becker, S. and Hinton, G. (1993). Learning mixture models of spatial coherence. *Neural Computation*, 5:267–277.

Becker, S. and Plumbley, M. (1996). Unsupervised neural network learning procedures for feature extraction and classification. *Journal of Applied Intelligence*, 6:1–21.

Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 19(7):711–720.

Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Bell, A. and Sejnowski, T. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Berns, G., Dayan, P., and Sejnowski, T. (1993). A correlational model for the development of disparity selectivity in visual cortex that depends on prenatal and postnatal phases. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):8277–81.

Beymer, D. (1994). Face recognition under varying pose. In *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 756–61, Seattle, WA. IEEE Comput. Soc. Press, Los Alamitos, CA.

Beymer, D. and Poggio, T. (1996). Image representations for visual learning. *Science*, 272(5270):1905–9.

Beymer, D., Shashua, A., and Poggio, T. (1993). Example based image analysis and synthesis. AI Memo 1431, Massachusetts Institute of Technology.

Biederman, I. (1998). Neural and psychophysical analysis of object and face recognition. In Wechsler, H. ., Phillips, P., Bruce, V., Fogelman-Soulie, F., and Huang, T., editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag.

Blakemore, C. (1991). Sensitive and vulnerable periods in the development of the visual system. *Ciba Foundation Symposium*, 156:129–47.

Breiter, H., Etcoff, N., Whalen, P., Kennedy, W., Rauch, S., Buckner, R., Strauss, M., Hyman, S., and Rosen, B. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17(5):875–87.

Bruce, V. (1988). *Recognising Faces*. Lawrence Erlbaum Assoc., London.

Bruce, V. (1998). Human face perception and identification. In Wechsler, H. ., Phillips, P., Bruce, V., Fogelman-Soulie, F., and Huang, T., editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag.

Brunelli, R. and Poggio, T. (1993). Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052.

Bulthoff, H., Edelman, S., and Tarr, M. (1995). How are three-dimensional objects represented in the brain. *Cerebral Cortex*, 3:247–260.

Callaway, E. and Katz, L. (1991). Effects of binocular deprivation on the development of clustered horizontal connections in cat striate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 88(3):745–9.

Camras, L. (1977). Facial expressions used by children in conflict situations. *Child Development*, 48:1431–35.

Cardoso, J.-F. and Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–30.

Carmignoto, G. and Vicini, S. (1992). Activity dependent decrease in nmda receptor responses during development of the visual cortex. *Science*, 258(5804):1007–11.

Chellappa, R. (1998). Discriminant analysis for face recognition. In Wechsler, H., Phillips, P., Bruce, V., Fogelman-Soulie, F., and Huang, T., editors, *Face Recognition: From Theory to Applications. NATO ASI Series F.* Springer-Verlag.

Cohn, J., Zlochower, A., Lien, J., Wu, Y.-T., and Kanade, T. (1999). Automated face coding: A computer-vision based method of facial expression analysis. *Psychophysiology*, 35(1):35–43.

Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, 36:287–314.

Constantine-Paton, M., Cline, H., and Debinski, E. (1990). Patterend activity, synaptic convergence, and the nmda receptor in developing visual pathways. *Annual Review of Neuroscience*, 13:129–154.

Cooper, E. (1998). Unpublished research, as cited by I. Biederman, (1998), in H . Wechsler, P.J. Phillips, V. Bruce, F. Fogelman-Soulie, T. Huang, (Eds.), Face Recognition: From Theory to Applications. Springer-Verlag.

Cottrell, G. (1990). Extracting features from faces using compression networks: Face, identity, emotion, and gender recognition using holons. Connectionist Models Summer School.

Cottrell, G. and Fleming, M. (1990). Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference*, pages 322–325, Dordrecht. Kluwer.

Cottrell, G. and Metcalfe, J. (1991). Face, gender and emotion recognition using holons. In Touretzky, D., editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA. Morgan Kaufmann.

Craig, K., Hyde, S., and Patrick, C. (1991). Genuine, suppressed, and faked facial behavior during exacerbation of chronic low back pain. *Pain*, 46:161–172.

Daugman, J. (1988). Complete discrete 2d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:1169–1179.

Davidson, R., Ekman, P., Saron, C., Senulis, J., and Friesen, W. (1990). Emotional expression and brain physiology i. approach/withdrawal and cerebral asymmetry. *Journal of Personality and Social Psychology*, 58:330–341.

Dayan, P., Hinton, G., Neal, R., and Zemel, R. (1995). The helmholtz machine. *Neural Computation*, 7(5):889–904.

Debinski, E., Cline, H., and Constantine-Paton, M. (1990). Activity-dependent tuning and the nmda receptor. *Journal of Neurobiology*, 21(1):18–32.

DeValois, R. and DeValois, K. (1988). *Spatial Vision.* Oxford Press.

Donato, G., Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989.

Ekman, P. (1982a). *Emotion in the Human Face, Second Edition.* Cambridge University Press, New York.

Ekman, P. (1982b). Methods for measuring facial action. In Scherer, K. and Ekman, P., editors, *Handbook of Methods in Nonverbal Behavior Research*, pages 45–135. Cambridge University Press, New York.

Ekman, P. (1985). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton, New York, 1st edition.

Ekman, P. (1991). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton, New York, 2nd edition.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, CA.

Ekman, P., Friesen, W., and O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, 545:414 – 420.

Ekman, P., Huang, T., Sejnowski, T., and Hager, J. (1992). Final report to NSF of the planning workshop on facial expression understanding. Available from UCSF, HIL-0984, San Francisco, CA 94143.

Ekman, P., Levenson, R., and Friesen, W. (1983). Autonomic nervous system activity distinguishes between emotions. *Science*, 221:1208–1210.

Ekman, P. and Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, 30:527–554.

Ekman, P. and Rosenberg, E. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York.

Essa, I. and Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–63.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America, A*, 4:2379–94.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Fisher, R. (1936). The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Fox, K., Daw, N., Sato, H., and Czepita, D. (1992). The effect of visual experience on the development of nmda receptor synaptic transmission in kitten visual cortex. *Journal of Neuroscience*, 13:155–69.

Freeman, W. (1994). Characterization of state transitions in spatially distributed, chaotic, nonlinear, dynamical systems in cerebral cortex. *Integrative Physiological and Behavioral Science*, 29(3):294–306.

Gallant, J., Connor, C., and Van Essen, D. (1994). Responses of visual cortical neurons in a monkey freely viewing natural scenes. In *Society for Neuroscience Abstracts*, volume 20, page 838.

Gardner, E. (1988). The space of interactions in neural network models. *Journal of Physics A: Math. Gen.*, 21:257–270.

Gibson, J. (1986). *The Ecological Approach to Visual Perception.* Lawrence Erlbaum, New Jersey.

Golomb, B., Lawrence, D., and Sejnowski, T. (1991). Sexnet: A neural network identifies sex from human faces. In Lippman, R., Moody, J., and Touretzky, D., editors, *Advances in Neural Information Processing Systems*, volume 3, pages 572–577. Morgan-Kaufmann, San Mateo, CA.

Gray, M., Movellan, J., and Sejnowski, T. (1997). A comparison of local versus global image decomposition for visual speechreading. In *Proceedings of the 4th Joint Symposium on Neural Computation*, pages 92–98. Institute for Neural Computation, La Jolla, CA, 92093-0523.

Griniasty, M., Tsodyks, M., and Amit, D. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 5:1–17.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part 1. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134.

Gu, Q. and Singer, W. (1993). Effects of intracortical infusion of anticholinergic drugs on neuronal plasticity in kitten striate cortex. *European Journal of Neuroscience*, 5(5):475–85.

Hager, J. and Ekman, P. (1995). The essential behavioral science of the face and gesture that computer scientists need to know. In Bichsel, M., editor, *Proceedings of the International Workshop on Automatic Face- and Gesture-*

Recognition, pages 7–11. Available from University of Zurich, Department of Computer Science, Winterhurrerstrasse 190, CH-8057.

Hallinan, P. (1995). *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions.* PhD thesis, Harvard University.

Hancock, P. (2000). Alternative representations for faces. British Psychological Society, Cognitive Section, University of Essex, September 6-8.

Hancock, P., Burton, A., and Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory and Cognition*, 24:26–40.

Harris, W. and Holt, C. (1990). Early events in the embryogenesis of the vertibrate visual system: Cellular determination and path finding. *Annual Review of Neuroscience*, 13:155–169.

Hasselmo, M., Rolls, E., Baylis, G., and Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 75(2):417–29.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation.* MacMillan, New Jersey.

Hebb, D. (1949). *The organization of Behavior*. Wiley, New York.

Heeger, D. (1991). Nonlinear model of neural responses in cat visual cortex. In Landy, M. and Movshon, J., editors, *Computational Models of Visual Processing*, pages 119–133. MIT Press, Cambridge, MA.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197.

Heller, M. and Haynal, V. (1994). The faces of suicidal depression (translation). les visages de la depression de suicide. *Kahiers Psychiatriques Genevois (Medecine et Hygiene Editors)*, 16:107–117.

Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–61.

Hinton, G. and Sejnowski, T. (1999). *Unsupervised learning: Foundations of Neural Computation.* MIT Press.

Hinton, G. and Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–75.

Hinton, G. and Zemel, R. (1994). Autoencoders, minimum description length, and helmholtz free energy. In Cowan, J., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 3–10, San Francisco, CA. Morgan Kaufmann.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, 79:2554–2558.

Horn, B. and Schunk, B. (1981). Determining optical flow. *Artificial Intelligence*, 17:185 – 203.

Hubel, D., Wiesel, T., and LeVay, S. (1977). Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical transactions of the Royal Society of London (Biol).*, 278:377–409.

Jain, R., Kasturi, R., and Schunck, B. (1995). *Machine Vision.* McGraw-Hill, New York.

Jenkins, W., Merzenich, M., and Recanzone, G. (1990). Neocortical representational dynamics in adult primates: implications for neuropsychology. *Neuropsychologia*, 28(6):573–84.

Jones, J. and Palmer, L. (1987). An evaluation of the two dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58:1233–1258.

Kaas, J. (1991). Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscience*, 14:137–67.

Kanade, T. (1973). Picture processing system by computer complex and recognition of human faces. Department of Information Science Technical Report November, Kyoto University.

Kanade, T. (1977). *Computer recognition of human faces.* Birkhauser Verlag, Basel and Stuttgart.

Kanfer, S. (1997). *Serious business : the art and commerce of animation in America from Betty Boop to Toy Story.* Scribner, New York.

Kinomura, S., Kawashima, R., Yamada, K., Ono, S., Itoh, M., Yoshioka, S., Yamaguchi, T., Matsui, H., Miyazawa, H., Itoh, H., and et al. (1994). Functional anatomy of taste perception in the human brain studied with positron emission tomography. *Brain Research*, 659(1-2):263–6.

Knill, D. and Richards, W. (1996). *Perception as Bayesian inference.* Cambridge University Press, New York.

Kohonen, T., Oja, E., and Lehtio, P. (1981). Storage and processing of information in distributed associative memory systems. In Hinton, G. and Anderson, J., editors, *Parallel Models of Associative Memory*, pages 49–81. Erlbaum, Hillsdale.

Lades, M., Vorbrüggen, J., Buhmann, J., Lange, J., Konen, W., von der Malsburg, C., and Würtz, R. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

Lander, K. and Bruce, V. (1997). The role of movement in the recognition of famous faces. Poster presentation, NATO ASI on Face Recognition: From Theory to Applications, Stirling, Scotland. Submitted for journal publication.

Lanitis, A., Taylor, C., and Cootes, T. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, 36:910–912.

Lee, D. and Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Lee, T.-W. (1998). *Independent Component Analysis: Theory and Applications.* Kluwer Academic Publishers.

Lee, T.-W., Girolami, M., and Sejnowski, T. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–41.

Lee, T.-W., Koehler, B., and Orglmeister, R. (1997). Blind source separation of nonlinear mixing models. In *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*, pages 406–415, Florida.

Lewicki, M. and Olshausen, B. (1998). Inferring sparse, overcomplete image codes using an efficient coding framework. In Jordan, M., editor, *Advances in Neural Information Processing Systems*, volume 10, San Mateo. Morgan Kaufmann.

Lewicki, M. and Olshausen, B. (1999). Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587–601.

Lewicki, M. and Sejnowski, T. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–65.

Li, H., Roivainen, P., and Forchheimer, R. (1993). 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555.

Lien, J., Kanade, T., J.F., C., and Li, C. (2000). Detection, tracking, and classification of action units infacial expression. *Journal of Robotics and Autonomous Systems*, 31(3):131–46.

Lin, J., Grier, D., and Cowan, J. (1997). Source separation and density estimation by faithful equivariant som. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 536–541, Cambridge, MA. MIT Press.

Linsker, R. (1986). From basis network principles to neural architecture (3 paper series). *Proceedings of the National Academy of Science, USA*, 83:7508–7512, 8390–8394, 8779–8783.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Logothetis, N. and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288.

MacKay, D. J. C. (1996). Hyperparameters: Optimise or interate out? In Heidbreder, G., editor, *Maximum entropy and Bayesian methods, Santa Barbara 1993*. Kluwer, Dordrecht.

Macleod, D. and von der Twer, T. (1996). Optimal nonlinear codes. Technical Report 28/96, Universität Bielefeld, Zentrum für interdisziplinäre Froschung.

Makeig, S., Bell, A., Jung, T.-P., and Sejnowski, T. (1996). Independent component analysis of electroencephalographic data. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, Cambridge, MA. MIT Press.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions E*, 74(10):3474–3483.

Mastronarde, D. (1989). Correlated firing of retinal ganglion cells. *Trends in Neuroscience*, 12(2):75–80.

McKelvie, S. (1995). Emotional expression in upside-down faces: Evidence for configurational and componential processing. *British Journal of Social Psychology*, 34(3):325–334.

McKeown, M., Makeig, S., Brown, G., Jung, T.-P., Kindermann, S., Bell, A., and Sejnowski, T. (1998). Analysis of fmri by decomposition into independent spatial components. *Human Brain Mapping*, 6(3):160–88.

Meister, M., Wong, R., Baylor, D., and Shatz, C. (1991). Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–43.

Miller, K., Keller, J., and Stryker, M. (1989). Ocular dominance column development: analysis and simulation. *Science*, 245(4918):605–15.

Millward, R. and O'Toole, A. (1986). Recognition memory transfer between spatial frequency analyzed faces. In Ellis, H., Jeeves, M., Newcombe, F., and Young, A., editors, *Aspects of Face Processing*, pages 34–44. Nijhoff, Dodrecht.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(27):817–820.

Montague, P., Gally, J., and Edelman, G. (1991). Spatial signaling in the development and function of neural connections. *Cerebral Cortex*, 1:199–220.

Morris, J., Frith, C., Perrett, D., Rowland, D., Young, A., Calder, A., and Dolan, R. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, 383(6603):812–5.

Movellan, J. (1995). Visual speech recognition with stochastic networks. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge, MA.

MPEG Video and SNHC (1998). Study of cd 14496-2 (visual). MPEG-4 Visual Committee Draft, ISO/IEC/JTC1/SC29/WG11/N2072.

Mumford, D. (1996). Pattern theory: a unifying perspective. In Knill, D. and Richards, W., editors, *Perception as Bayesian Inference*. Cambridge University Press.

Nadal, J.-P. and Parga, N. (1994). Non-linear neurons in the low noise limit: a factorial code maximizes information trans fer. *Network*, 5:565–581.

Nowlan, S. (1990). Maximum likelihood competitive learning. In Touretzky, D., editor, *Neural Information Processing Systems*, volume 2, pages 574–582, San Mateo, CA. Morgan-Kaufmann.

Obermayer, K. and Blasdel, G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience*, 13(10):4114–29.

Obermayer, K., Blasdel, G., and Schulten, K. (1992). Statistical-mechanical analysis of self-organization and pattern formation during development of visual maps. *Physical Review A.*, 45(10):7568–89.

Oja, E. (1982). A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, 15:267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68.

Olshausen, B. and Field, D. (1996a). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Olshausen, B. and Field, D. (1996b). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–340.

Oppenheim, A. and Lim, J. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69:529–541.

O'Reilly, R. and Johnson, M. (1994). Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, 6:357–389.

O'Toole, A., Abdi, H., Deffenbacher, K., and Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, 10(3):405–411.

O'Toole, A., Deffenbacher, K., Valentin, D., and Abdi, H. (1994). Structural aspects of face recognition and the other race effect. *Memory and Cognition*, 22(2):208–224.

Padgett, C. and Cottrell, G. (1997). Representing face images for emotion classification. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA. MIT Press.

Palm, G. (1980). On associative memory. *Biological Cybernetics*, 36:19–31.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes.* McGraw-Hill, New York.

Parga, N. and Rolls, E. (1998). Transform invariant recognition by association in a recurrent network. *Neural Computation*, 10(6):1507–25.

Pearlmutter, B. and Hinton, G. (1986). G-maximization: An unsupervised learning procedure for discovering regularities. In Denker, J., editor, *Neural Net-*

works for Computing: American Institute of Physics Conference Proceedings, volume 151, pages 333–338.

Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ica. In Mozer, Jordan, and Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press,.

Penev, P. (2001). Redundancy and dimensionality reduction in sparse-distributed representations of natural objects in terms of their local features. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*. MIT Press.

Penev, P. and Atick, J. (1996). Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500.

Pentland, A., Moghaddam, B., and Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.*

Perrett, D., Mistlin, A., and Chitty, A. (1989). Visual neurones responsive to faces. *Trends in Neuroscience*, 10:358–364.

Peruš, M. (2000). From neural to quantum associative networks: a new quantum "algorithm". In Dubois, D., editor, *AIP Conference Proceedings on Computing Anticipatory Systems*, number 517, pages 289–95.

Peruš, M. (2001). A synthesis of the pibram holonomic theory of vision with quantum associative nets after pre-processing using i.c.a. and other computational models. *International Journal of Computing Anticipatory Systems*, in press.

Phillips, M., Young, A., Senior, C., Brammer, M.and Andrews, C., Calder, A., Bullmore, E., Perrett, D., Rowland, D., Williams, S., Gray, A., and David, A. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature*, 389:495–498.

Phillips, P., Wechsler, H., Juang, J., and Rauss, P. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306.

Picard, R. (1997). *Affective Computing*. MIT Press.

Pike, G., Kemp, R., Towell, N., and Phillips, K. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, 4(4):409–437.

Piotrowski, L. and Campbell, F. (1982). A demonstration of the visual importance and flexibility of spatial-frequency, amplitude, and phase. *Perception*, 11:337–346.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266.

Pollen, D. and Ronner, S. (1981). Phase relationship between adjacent simple cells in the visula cortex. *Science*, 212:1409–1411.

Pratt, W. (1978). *Digital Image Processing.* Wiley, New York.

Rhodes, P. (1992). The long open time of the nmda channel facilitates the self-organization of invariant object responses in cortex. In *Society for Neuroscience Abstracts*, volume 18, page 740.

Rison, R. and Stanton, P. (1995). Long-term potentiation and n-methyl-d-aspartate receptors: foundations of memory and neurologic disease. *Neuroscience and Biobehavioral Reviews*, 19(4):533–52.

Rolls, E. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):11–20.

Rolls, E. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research*, 66:177–185.

Rolls, E., Baylis, G., and Hasselmo, M. (1987). The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vision Research*, 27(3):311–26.

Rosenblum, M., Yacoob, Y., and Davis, L. (1996). Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138.

Rumelhart, D. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9:75–112.

Rydfalk, M. (1987). *CANDIDE: A parametrized face.* PhD thesis, Linkoping University, Department of Electrical Engineering.

Sanger, T. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473.

Schraudolph, N. and Sejnowski, T. (1992). Competitive anti-hebbian learning of invariants. In Moody, J., Hanson, S., and Lippman, R., editors, *Advances in Neural information Processing Systems*, volume 4, pages 1017–1024, San Francisco. Morgan Kaufmann.

Shannon, C. and Weaver, W. (1949). *The Mathematic Theory of Communication.* University of Illinois Press, Urbana, IL.

Shashua, A. (1992). *Geometry and Photometry in 3D Visual Recognition.* PhD thesis, Massachusetts Institute of Technology.

Shepard, R. and Cooper, L. (1982). *Mental Images and their Transformations.* MIT Press, Cambridge, MA.

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA.

Singer, W. (1990). The formation of cooperative cell assempblies in the visual cortex. *Journal of Experimental Biology*, 153:177–197.

Singh, A. (1991). *Optic Flow Computation.* IEEE Computer Society Press, Los Alamitos, CA.

Sone, J., Porrill, J., Buchel, C., and Friston, K. (1999). Spatial, temporal, and spatiotemporal independent component analysis of fmri data. In *18th Leeds Statistical Research Workshop on Spatio-temporal modelling anad its applications*.

Stone, J. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–92.

Stone, J. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, 38(7):947–951.

Stone, J. and Porrill, J. (1998). Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, University of Sheffield, Department of Psychology.

Stryker, M. (1991a). Activity-dependent reorganization of afferents in the developing mamalian visual system. In Lam, D. and Schatz, C., editors, *Development of the Visual System*, pages 267–287. MIT Press, Cambridge, MA.

Stryker, M. (1991b). Temporal associations. *Nature*, 354(14):108–109.

Stryker, M. and Harris, W. (1986). Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *Journal of Neuroscience*, 6:2117–2133.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, 262:685–688.

Teh, Y. and Hinton, G. (2001). Rate-coded restricted boltzmann machines for face recognition. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*. MIT Press.

Terzopoulus, D. and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579.

Tian, Y., Kanade, T., and Cohn, J. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2).

Tranel, D., Damasio, A., and Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, 38(5):690–696.

Tsodyks, M. and Feigel'man, M. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6(2):101–105.

Tukey, J. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29:614.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Turrigiano, G., Leslie, K., Desai, N., Rutherford, L., and Nelson, S. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–6.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006.

Valentin, D., Abdi, H., O'Toole, A., and Cottrell, G. (1994). Connectionist models of face processing: a survey. *Pattern Recognition*, 27(9):1209–30.

Vetter, T. and Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 15(6):733–741.

Vetter, T. and Troje, N. (1997). Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, 14(9):2152–61.

Wachtler, T., Lee, T.-W., and Sejnowski, T. (2001). The chromatic structure of natural scenes. *Journal of the Optical Socitey of America, A*, 18(1):65–77.

Wallbott, H. (1992). Effects of distortion of spatial and temporal resolution of video stimuli on emotion attributions. *Journal of Nonverbal Behavior*, 15(6):5–20.

Wallis, G. and Baddeley, P. (1997). Optimal, unsupervised learning in invariant object recognition. *Neural Computation*, 9(4):883–94.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology (Oxford)*, 51(2):167–194.

Weinshall, D. and Edelman, S. (1991). A self-organizing multiple view representation of 3d objects. *Biological Cybernetics*.

Yacoob, Y. and Davis, L. (1994). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642.

Yang, H., Amari, S.-I., and Cichocki, A. (1998). nformation-theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–3000.

Yaxley, S., Rolls, E., and Sienkiewicz, Z. (1988). The responsiveness of neurons in the insular gustatory cortex of the macaque monkey is independent of hunger. *Physiology and Behavior*, 42(3):223–9.

Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 56:1327–1331.

Zajonc, R. (1984). The interaction of affect and cognition. In Scherer, K. and Ekman, P., editors, *Approaches to Emotion*, pages 239–246. Lawrence Erlbaum, Hillsdale, NJ.

Zemel, R. and Hinton, G. (1991). Discovering viewpoint invariant objects that characterize objects. In Lipmann, R., Moody, J., and Touretzky, D., editors, *Advances in Neural Information Processing Systems*, volume 3, pages 299–305, San Francisco. Morgan Kaufmann.

Zhang, J., Yan, Y., and Lades, M. (1997). Face recognition: Eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, 85(9):1423–1435.

Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings, Third IEEE International Conference on Automatic Faceand Gesture Recognition*, pages 454–459. IEEE Computer Society.

Index

Affordances, 5
Attractor network, 31, 132, 138, 148
Autoassociative memory, 35
Basin of attraction, 31, 132–133, 137, 140, 147
Bayesian inference models, 6
  and minimum description length, 8
Coding strategies, 3, 41
  and information maximization, 15
  independence, 9
  redundancy reduction, 8
Competitive learning rule, 132, 134
Complex cells, 29
Deceit detection, 78–80
Desirable filters, 3, 8, 41
Efficient codes, 8
  advantages, 9
  and principal component analysis, 12
  and redundancy reduction, 11
Eigenfaces, 34, 48, 56, 62–63
  and facial expression analysis, 73, 85, 108
Eigenfeatures, 37, 117
Elastic matching algorithm, 37
Entropy, 10
  and independent component analysis, 20
  joint, 10
FERET face database, 44
Face cells, 35, 147
Face recognition
  algorithms, 33
  and Fisher's linear discriminants, 36
  and Gabor filters, 37
  and local feature analysis, 36
  and principal component analysis, 34
  eigenfaces, 48, 56, 62
  independent component analysis, 48, 56, 62
  principal component analysis, 48, 56, 62
Facial action database, 84
Facial expression recognition, 70, 74
  algorithms, 70–73
  by Fisher's linear discriminants, 112
  by Gabor filters, 119
  by eigenfaces, 85, 108
  by hidden Markov models, 127
  by independent component analysis, 114
  by local PCA, 117
  by local feature analysis, 109
  by optic flow, 88, 105
  by wrinkle measurement, 87
  neural substrates, 98
  posed vs. spontaneous expressions, 79
  processing speed, 96
Features versus templates, 72, 98
Fisher's linear discriminants, 36, 112
Fixed points, 31, 138, 142
GMAX algorithm, 15
Gabor filters, 37, 119
  and face recognition, 37
  and facial expression recognition, 119
  and independent component analysis, 126
  and sparseness, 23
Generative models, 6
Hebbian learning, 13, 130, 132, 137, 142
  and NMDA channel, 24
  and information maximization, 14
  and invariance, 29, 32
  and principal component analysis, 14, 148
  and visual development, 24, 26–28
Helmholtz machine, 7
Hidden Markov models, 127
High-order dependencies, 18
Holons, 34, 63
Hopfield network, 132, 139, 146
IMAX learning rule, 16
  and transformation invariance, 17
Image basis, 126, 156
Image registration, 103
Image synthesis model, 45, 54
Independence
  and sparse coding, 22

versus decorrelation, 17
Independent componene analysis
and human perception, 67
Independent component analysis, 18
and Gabor filters, 126
and entropy maximization, 20
and face recognition, 48, 56, 62
and facial expression recognition, 114
and maximum likelihood, 155
and natural scenes, 23
and sparseness, 61
and spectral sensitivity, 24
basis images, 48
face representation, 47–48, 56
factorial code, 53
learning rule, 21, 42
nonlinear, 155
number of sources, 154
spatio-temporal, 154
vs. principal component analysis, 18–19, 64
website for Matlab code, 42
Inferior temporal lobe, 32, 129–131, 147
Information maximization
ICA learning rule, 21
and Hebbian learning, 14
and coding strategies, 15
and independent component analysis, 18
and minimum entropy coding, 21
Information, 9
Invariance, 132
and NMDA channel, 132
and hebbian learning, 32
in the visual system, 32–33, 130
learning rules, 16–17, 29–31, 132–133, 143, 148
Kurtosis, 20, 61
Large monopolar cells (LMC), 11
Likelihood, 6
Local feature analysis, 36, 65, 109
MPEG-4, 77
Maximum likelihood competitive learning, 6
Maximum likelihood models, 6
and wake-sleep algorithm, 8
Mental rotation, 149
Minimum description length, 7
and bayesian inference models, 8
Minimum entropy coding, 10
and information maximization, 21
and sparseness, 22
Mixture of gaussians model, 6
Mutual information, 14
and invariance, 16
and stereo depth, 17
in image representations, 59
NMDA channel
and Hebbian learning, 24
and learning invariances, 32, 132
and visual development, 25
Natural gradient, 21
Natural scenes
amplitude spectrum, 11
and independent component analysis, 23
and sparseness, 23
Nearest neighbor algorithm, 48
Non-negative matrix factorization, 65
Ocular dominance, 24–25, 27–28
Optic flow
correlation-based, 105
gradient based, 88
Optimal gain control, 11
and independent component analysis, 19–20
and spectral sensitivity, 11
Orientation columns, 24–26, 28
Other race effect, 35
Phase, 18
Pleistochrome, 12
Pose tuning, 32, 130, 140
Principal component analysis, 12
and Gaussian models, 13
and Hebbian learning, 14, 148
and autoassociative memory, 35
and face recognition, 34, 48, 56, 62
and facial expression recognition, 73, 85, 108
and human perception, 35
basis, 13, 48
local principal component analysis, 37, 117
spatio-temporal, 137, 148
weights, 13
Radial basis functions, 71, 149
Receiver-operator-characteristic (ROC), 140
Receptive field, 15
Recurrent network, 132
Redundancy and structure, 8
Redundancy reduction, 8
and efficient coding, 11
in the visual system, 11, 15
and spectral sensitivity, 11
learning rules, 15
Retinal transfer function, 11
Self-organization of the visual system, 24
models, 26
ocular dominance, 25
orientation columns, 25
orientation specificity, 25
Simple cells, 23
Source separation, 21
Sparseness
advantages, 22
and Gabor filters, 23
and independence, 22
and minimum entropy coding, 22
and responses of visual neurons, 146
and storage capacity, 144, 154
of image representations, 60, 65
Spectral sensitivity

and optimal gain control, 11
Sphering, 42
Storage capacity, 144
Sub-Gaussian, 22, 155
Super-Gaussian, 20, 155
Temporal association, 29, 32, 130–131
Temporal filter, 132, 138, 140
The Facial Action Coding System, 75, 77
Unsupervised learning, 5
and visual development, 24
Visual cortical cells, 134
Wake-sleep algorithm, 7–8
and maximum likelihood models, 8
Whitening filter, 11, 36, 42
Wrinkle measurement, 87

About the Author

Marian Stewart Bartlett is a postdoctoral researcher at the Institute for Neural Computation, University of California, San Diego. She received her Bachelor's degree in Mathematics and Computer Science from Middlebury College in 1988, and her Ph.D. in Cognitive Science and Psychology from the University of California, San Diego in 1998. Her dissertation work was conducted with Terrence Sejnowski in the Computational Neurobiology Laboratory at the Salk Institute. Her interests include approaches to image analysis through unsupervised learning, with a focus on face recognition and expression analysis. She is presently exploring probabilistic dynamical models and their application to video analysis at the University of California, San Diego. Dr. Bartlett has also studied perceptual and cognitive processes with V.S. Ramachandran at the University of California San Diego, the Cognitive Neuroscience Section of the National Institutes of Health, the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology, and the Brain and Perception Laboratory at the University of Bristol, England. Dr Bartlett is married to Nigel Bartlett, and they have a son Paul Stewart Bartlett, born in 1999.