

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Face Image Analysis by Unsupervised Learning and Redundancy Reduction

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science and Psychology

by

Marian Stewart Bartlett

Committee in charge:

Terrence J. Sejnowski, Dissertation Adviser
Donald I. A. Macleod, Committee Chair
Gary Cottrell
Karen Dobkins
Harold Pashler
Martin Sereno

1998

Copyright
Marian Stewart Bartlett, 1998
All rights reserved.

The dissertation of Marian Stewart Bartlett is approved, and it is acceptable in quality and form for publication on microfilm:

~~Harrison W. Colwell~~

T. Synovick

Hal Pailin

Karen Hoff

A. J.

Don MacLeod

Chair

University of California, San Diego

1998

To my parents

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	xiii
Acknowledgements	xiv
Vita and Publications	xv
Abstract of the Dissertation	xvii
1 Introduction	1
1.1 Summary	1
1.2 Unsupervised Learning in Object Representations	3
1.2.1 Generative models	3
1.2.2 Redundancy reduction as an organizational principle	6
1.2.3 Principal component analysis	8
1.2.4 Hebbian learning	9
1.2.5 Learning rules for explicit discovery of statistical dependencies	11
1.2.6 High-order statistical dependencies	13
1.2.7 Self-organization of the visual system through correlation sensitive mechanisms	16
1.2.8 Learning invariances from temporal dependencies in the input	19
1.3 Computational Algorithms for Recognizing Faces in Images	22
2 Independent Component Representations for Face Recognition	27
2.1 Abstract	27
2.2 Introduction	28
2.3 Independent Component Analysis (ICA)	29
2.4 Independent Component Representations of Face Images	31
2.4.1 Statistically independent basis images	31
2.4.2 Independence in face space versus pixel space	34
2.4.3 A factorial face code	36
2.5 Face Recognition Performance	37
2.5.1 Independent basis architecture	38

2.5.2	Factorial code architecture	42
2.6	Examination of the ICA Representations	46
2.6.1	Mutual information	46
2.6.2	Sparseness	48
2.6.3	Combined ICA recognition system	49
2.7	Discussion	50
3	Measuring Facial Expressions by Computer Image Analysis	55
3.1	Abstract	55
3.2	Introduction	55
3.2.1	Measurement of facial signals	56
3.2.2	Analysis of facial signals by computer	57
3.3	Automating the Facial Action Coding System (FACS)	59
3.3.1	Methods	60
3.3.2	Results	66
3.4	Discussion	72
4	Classifying Facial Actions	76
4.1	Abstract	76
4.2	Introduction	76
4.2.1	The Facial Action Coding System	77
4.2.2	Analysis of facial signals by computer	78
4.3	Overview of Approach	84
4.4	Image Database	85
4.5	Optic Flow Analysis	85
4.5.1	Gradient-based optic flow	85
4.5.2	Correlation-based optic flow	87
4.5.3	Local smoothing	88
4.6	Holistic Analysis	89
4.6.1	Principal Component Analysis: “EigenActions”	89
4.6.2	“FisherActions”	89
4.6.3	Independent Component Analysis	92
4.6.4	Local Feature Analysis (LFA)	93
4.7	Local Representations	96
4.7.1	Gaussian kernel	96
4.7.2	Local PCA: Random patches	96
4.7.3	Local PCA: Fixed patches	97
4.7.4	Gabor wavelet representation	97
4.7.5	PCA jets	99
4.8	Human Subjects	99
4.8.1	Naive subjects	99
4.8.2	Expert coders	101
4.9	Performance	101

4.9.1	Classification procedures	101
4.9.2	Optic flow analysis	102
4.9.3	Holistic spatial analysis	102
4.9.4	Local analysis	103
4.10	Discussion	104
4.11	Conclusions	107
5	Learning Viewpoint Invariant Face Representations from Visual Experience in an Attractor Network	111
5.1	Abstract	111
5.2	Introduction	111
5.3	Simulation	115
5.3.1	Model architecture	115
5.3.2	Competitive Hebbian learning of temporal relationships	116
5.3.3	Temporal association in an attractor network	118
5.3.4	Simulation results	120
5.4	Discussion	128
6	Conclusions and Future Directions	132
	References	136

LIST OF FIGURES

1.1	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).	6
1.2	Example 2-D data distribution and the corresponding principal component and independent component axes. Figure inspired by Lewicki & Sejnowski (submitted).	15
2.1	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 <i>Neural Computation</i> , copyright 1995, MIT Press.	30
2.2	Example 2-D data distribution and corresponding principal component and independent component axes. Figure inspired by Lewicki & Sejnowski (submitted).	31
2.3	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 .	32
2.4	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} .	33
2.5	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.	35

2.6	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	36
2.7	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}	37
2.8	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.	37
2.9	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 2.3). Independent components are ordered by the class discriminability ratio, r (Equation 2.8).	39
2.10	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.	40
2.11	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.	41
2.12	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.	43
2.13	Basis images for the ICA factorial representation (columns of $A = W_I^{-1}$) obtained with Architecture 2. (See Figure 2.6).	44
2.14	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.	45
2.15	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.	46

2.16	Pairwise mutual information. Top: 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. Bottom: 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.	47
2.17	Kurtosis (sparseness) of ICA and PCA representations.	49
2.18	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.	50
2.19	Face recognition performance of the combined ICA classifier, compared to the individual classifiers for ICA1 and ICA2, and PCA.	51
3.1	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.	61
3.2	Figure 1 test	62
3.3	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.	63
3.4	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 i th pixel of the segment. Pixel differences approximate the derivative (Jain, Kasturi, & Schunk, 1995). d) P measured at image location A for one subject performing each of the six actions.	64
3.5	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.	65
3.6	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.	68
4.1	Example action sequences from the database. The example sequences show two subjects demonstrating AU1 starting from a null expression and ending with a high magnitude example of the action. Frame 2 is a low magnitude example of the action, frames 3 and 4 are medium magnitude examples, and frames 5 and 6 are high magnitude.	80

4.2	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, Action Unit name, and number of subjects imaged.	86
4.3	Optic flow for AU1 extracted using local velocity information extracted by the correlation-based technique, with no spatial smoothing.	88
4.4	First 8 principal components of the difference images for the upper face actions (top), and lower face actions (bottom). Components are ordered left to right, top to bottom.	90
4.5	TOP: Image synthesis model for the ICA representation. BOTTOM: The ICA representation.	93
4.6	Example independent component images of the upper face (top) and lower face actions (bottom).	94
4.7	An original δ -image on the left, with its corresponding LFA representation $O(x)$ on the right.	95
4.8	The first 155 points selected by the sparsification algorithm superimposed to the mean images of the upper and lower face actions.	96
4.9	The first 10 principal components of 7750 patches extracted from random locations in the δ -images. Components are ordered left to right, top to bottom. . . .	97
4.10	Local principal components of 15 x 15 patches in fixed locations of the upper face δ -images. From top to bottom: Principal components 1-3.	98
4.11	Top: Original δ -image. Bottom two rows, from left to right: Gabor kernels (low and high frequency), the imaginary part and magnitude of the filtered image. . .	100
4.12	PCA Jets. Left: two kernels corresponding to low and high frequencies (patches size 49×49 and 9×9). Right: the result of the convolution with the δ -image of Figure 4.11.	100
5.1	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 <i>Nature</i> , copyright 1988, MacMillan Magazines, Ltd. . . .	113
5.2	Sample of the 100 images used in the simulation. Image set provided by David Beymer (1994).	115
5.3	Model architecture.	116

5.4	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.	121
5.5	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.	122
5.6	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).	124
5.7	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).	125
5.8	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 $\lambda = 0.5$	127

LIST OF TABLES

2.1	Image sets used for training and testing.	38
3.1	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.	70
3.2	Confusion Matrix for the automated classifiers.given action. Hol: Holistic, Mt: Motion, Ft: Feature, Hyb: Hybrid.	71
3.3	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***.	72
4.1	Best performance for each classifier. FLD: Fisher's linear discriminant. ICA: Independent component analysis. LFA: Local feature analysis.	105
4.2	Summary of automated facial expression analysis systems. These models were tested on different data sets with differing levels of difficulty. Classification categories were feigned expressions of emotion, except where otherwise indicated. ^a Classified facial actions ^b Classified images from Pictures of Facial Affect (Ekman & Friesen, 1976) in which trained actors performed the muscle contractions empirically associated with certain emotional states.	108
5.1	Contribution of the feedforward connections and the attractor network to view-point 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).	123
5.2	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 $\frac{F}{N} < 0.14$	128

ACKNOWLEDGEMENTS

My biggest debt of gratitude goes to Nigel for his love and support throughout this endeavor. It has been a great privilege to work with my thesis adviser, Terry Sejnowski, for the past five years. I have 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 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 thesis. I also thank my office-mate Michael Gray for sharing ideas, space, and experiences over more than five years of graduate school. Finally, I would like to thank my parents, to whom this thesis is dedicated, for always being supportive. The text of Chapter Two, in part, is a reprint of the material as it appears in *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, Vol. 3299. I was the primary researcher and author of this paper. The text of Chapter Three, in full, is a reprint of the material that has been accepted for publication in *Psychophysiology*. I was the primary researcher and author of this paper. The text of Chapter Four, in full, is a reprint of the material that has been submitted for publication in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. I was secondary researcher, primary writer, and project supervisor. The text of Chapter Five, in full, is a reprint of the material that is in press for *Network: Computation in Neural Systems*. I was primary researcher and author of this paper.

VITA

November 2, 1966	Born, Palo Alto, California
1988	B. A., <i>magna cum laude</i> , Mathematics and Computer Science, Middlebury College
1988–1989	Research Assistant, Massachusetts Institute of Technology
1989–1991	Research Assistant, National Institutes of Health
1994	M.A., Psychology, University of California San Diego
1998	Ph.D., Cognitive Science and Psychology, University of California San Diego

PUBLICATIONS

1. Bartlett, M.S., Hager, J.C., Ekman, P., and Sejnowski, T.J. (1998). Measuring facial expressions by computer image analysis. *Psychophysiology*, in press.
2. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., and Sejnowski, T.J. (submitted). Classifying Facial Actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
3. Bartlett, M. Stewart, and Sejnowski, T.J. (1998). Learning viewpoint invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems* 9(3) 399-417.
4. Bartlett, M. Stewart, Lades, H. Martin, and Sejnowski, T.J. (1998). Independent component representations for face recognition. *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Conference on Human Vision and Electronic Imaging III, San Jose, CA, January, 1998*, in press.
5. Bartlett, M. Stewart, and Sejnowski, T.J. (1998). Learning View point Invariant Face Representations from Visual Experience by Temporal Association. In H. Wechsler, P.J. Phillips, V. Bruce, S. Fogelman-Soulie, T. Huang (Eds.), *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag, in press.
6. Bartlett, M. Stewart, and Sejnowski, T. J. (1997). Viewpoint invariant face recognition using independent component analysis and attractor networks. In M. Mozer, M. Jordan, T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA. 817-823.
7. Bartlett, M. Stewart, and Sejnowski, T. J. (1997). Independent components of face images: A representation for face recognition. *Proceedings of the 4th Annual Joint Symposium on Neural Computation*, Pasadena, CA, May 17, 1997. Proceedings can be obtained from the Institute for Neural Computation, UCSD 0523, La Jolla, CA 92093.

8. Bartlett, M. Stewart, Viola, P. A., Sejnowski, T. J., Golomb, B.A., Larsen, J., Hager, J. C., and Ekman, P. (1996). Classifying facial action. In D. Touretzky, M. Mozer, M. Hasselmo, (Eds.) *Advances in Neural Information Processing Systems* 8, MIT Press, Cambridge, MA. p. 823-829.
9. Bartlett, M. Stewart, and Sejnowski, T.J. (1996). Unsupervised learning of invariant representations of faces through temporal association. In *Computational Neuroscience: International Review of Neurobiology Suppl. 1*. J.M. Bower, ed. Academic Press, San Diego, CA., 1996. p. 317-322.
10. Pascual-Leone, A, Grafman, J., Clark, K., Stewart M., Massaquoi, S., Lou JS., and Hallett M. (1993). Procedural learning in Parkinson's disease and cerebellar degeneration. *Annals of Neurology* 34(4); p.594-602.
11. Ramachandran, V. S., Rogers-Ramachandran, D. C., and Stewart, M. I. (1992) Perceptual Correlates of Massive Cortical Reorganization. *Science* 258; p. 1159-1160.
12. Grafman, J., Litvan, I., Massaquoi, S., Stewart, M., Sirigu, A., and Hallett, M. (1992). Cognitive planning deficit in patients with cerebellar atrophy. *Neurology* 42(8); p. 1493-1496.
13. Ramachandran, V.S., Stewart, M.I., and Rogers-Ramachandran, D.C. (1992). Perceptual Correlates of Massive Cortical Reorganization. *Neuroreport* 3(7), p. 583-586.
14. Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., and O'Connell, K.M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology, Human Perception and Performance*, 18(1); p. 34-49.
15. Newman, N.J., Wolfe, J.M., Stewart, M.I., and Lassell, S. (1991). Binocular visual function in patients with a history of monocular optic neuritis: Color and contrast sensitivity. *Clinical Vision Sciences* 6(2); p. 95-107.
16. Wolfe, J.M., Yu, K.P., Stewart, M.I., Shorter, A.D., and Cave, K.R. (1990). Limitations on the Parallel Guidance of Visual Search." *Journal of Experimental Psychology; Human Perception and Performance*, 16(4); p. 879-892.

ABSTRACT OF THE DISSERTATION

Face Image Analysis by Unsupervised Learning and Redundancy Reduction

by

Marian Stewart Bartlett

Doctor of Philosophy in Cognitive Science and Psychology
University of California San Diego, 1998

Terrence J. Sejnowski, Dissertation Adviser
Donald I. A. Macleod, Committee Chair

In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels. Representations such as "Eigenfaces" [197] and "Holons" [48] are based on Principal component analysis (PCA), which encodes the *correlational* structure of the input, but does not address high-order statistical dependencies such as relationships among three or more pixels. Independent component analysis (ICA) is a generalization of PCA which encodes the high-order dependencies in the input in addition to the correlations. Representations for face recognition were developed from the independent components of face images. The ICA representations were superior to PCA for recognizing faces across sessions and changes in expression.

ICA was compared to more than eight other image analysis methods on a task of recognizing facial expressions in a project to automate the Facial Action Coding System [62]. These methods included estimation of optical flow; representations based on the second-order statistics of the full face images such as Eigenfaces [47, 197] local feature analysis [156], and linear discriminant analysis [23]; and representations based on the outputs of local filters, such as a Gabor wavelet representations [50, 113] and local PCA [153]. The ICA and Gabor wavelet representations achieved the best performance of 96% for classifying 12 facial actions. Relationships between the independent component representation and the Gabor representation are discussed.

Temporal redundancy contains information for learning invariances. Different views of a face tend to appear in close temporal proximity as the person changes expression, pose, or moves through the environment. The final chapter 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 [72] 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.

Chapter 1

Introduction

1.1 Summary

Horace Barlow has argued that redundancy provides knowledge [8]. According to this theory, statistical redundancy contains information about the structure of the environment, and the important information for a perceptual system to detect is “suspicious coincidences”, new statistical regularities in the sensory input that differ from the environment to which the system 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 therefore constitute a “suspicious coincidence” [9]. 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.

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 [8]. Contrast gain control, which has been described in V1 [91], takes account of changes in the variance of the input signals. 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 [7] have argued for such decorrelation mechanisms as a general coding strategy for the visual system.

Some of the most successful algorithms for face recognition are based on learning mechanisms that are sensitive to the correlations in the face images. For example, representations such as “Eigenfaces” [197] “Holons” [48] and “Local Feature Analysis” [156] are data-driven face representations based on principal component analysis. Principal component analysis separates the correlations in the input, but does not address high order dependencies such as the relationships among three or more pixels. Edges are an example of a high-order dependence 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.

Independent component analysis [43] 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 [24] recently developed an algorithm for separating the statistically

independent components of a dataset. The algorithm is an unsupervised learning rule based on the principle of maximum information transfer in sigmoidal neurons [116]. The algorithm maximizes the mutual information between the input and the output of a transfer function by maximizing the joint entropy of the output. This produces statistically independent outputs under certain conditions. This algorithm has proven successful for separating randomly mixed auditory signals (the cocktail party problem), and has recently been applied to separating the sources of EEG signals [128], fMRI images [131], and has been applied to images of natural scenes to examine image filters that produce independent outputs [25].

Chapter 2 develops representations for face recognition based on statistically independent components of face images. ICA was performed on a set of face images by applying Bell and Sejnowski's information maximization algorithm [24]. Two ICA representations were developed, one which separated a set of independent images across spatial location, and a second which separated a set of independent coding features across face images. The ICA representations were compared to the "Eigenface" representation that is based on PCA. Recognition performance with the ICA representations was superior to "Eigenfaces" for recognizing faces across sessions, and changes in expression. A combined classifier that took account of the similarities within both ICA representations outperformed PCA for recognizing images collected within the same session as well.

Chapters 3 and 4 compared image representations for facial expression analysis, and demonstrate that representations derived from redundancy reduction on the graylevel face image ensemble are powerful for face image analysis. An independent component representation developed in Chapter 2 was compared to a number of other face image representation algorithms for analyzing facial expressions in a project to automate the Facial Action Coding System [62]. The Facial Action Coding System 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. These two chapters explored and compared techniques for automatically recognizing facial actions in sequences of images.

Chapter 3 compared holistic spatial analysis based on PCA, explicit measurement of features such as wrinkles, and estimation of motion flow fields for classifying facial actions. Performance of these systems was compared to naive and expert human subjects. The principal component representation, which extracted the second-order redundancies in the face images, gave better recognition performance than a set of hand-crafted feature measurements. The results also suggest that hand-crafted features plus holistic analysis such as PCA may be superior to either one alone, since their performances may be uncorrelated. A system combining the three representations performed as well as expert human subjects on this task.

Padgett & Cottrell [153] have shown that local filters can give better performance than holistic filters for classifying facial expressions. Chapter 4 compares over eight image representations including analysis of facial motion; holistic spatial analysis based on second-order image statistics such as principal component analysis, local feature analysis, and linear discriminant analysis; the independent component representation that made use of the high-order statistics as well; and representations based on the outputs of local filters, such as Gabor wavelets and local principal component analysis. Best performance was obtained using the Gabor wavelet repre-

sensation and the independent component representation, which both achieved 96% accuracy for classifying twelve facial actions. The results provide converging evidence for the importance of possessing local filters, high spatial frequencies, and statistical independence for classifying facial actions. Relationships between the independent component representation and the Gabor representation are discussed.

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 5 examines unsupervised learning of viewpoint invariant representations of faces through spatio-temporal redundancy reduction. This work explores the development of viewpoint invariant responses to faces from visual experience in a biological system. 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. 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. A set of simulations 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 [72] 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.

1.2 Unsupervised Learning in Object Representations

How can a perceptual system develop to recognize properties of its environment without being told which features it should analyze, or whether its identifications 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 [77]. Structure and information are contained in the external stimulus, and it is the task of the perceptual system to discover this structure. 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. (See [22] for a review of unsupervised learning.)

1.2.1 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. The model parameters can be considered to be network weights. 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 .

In this framework, the network is viewed as a probabilistic, generative model of the data. Let $\mathbf{x} = [x_1 \dots x_n]$ denote the observed data where the n samples are independent. The probability of the data given the model is given by

$$P(\mathbf{x}|\Phi) = \sum_i P(\mathbf{x}|\phi_i)P(\phi_i) \quad (1.1)$$

$$= \prod_j \sum_i P(x_j|\phi_i)P(\phi_i) \quad (1.2)$$

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 1.2. The log probability, or likelihood, is usually maximized in order to facilitate calculation of the partial derivatives of 1.2 with respect to each of the parameters. The model parameters are treated as network weights in an unsupervised learning framework. 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 [110]. The probability of the model given the data is given by

$$P(\Phi|\mathbf{x}) = \frac{P(\mathbf{x}|\Phi)P(\Phi)}{P(\mathbf{x})} \quad (1.3)$$

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.

Maximum likelihood competitive learning [142] is an extension of a mixture of Gaussians model. As in the mixture of Gaussians model, the posterior probability $p(x_j|\phi_i)$ is given by a Gaussian with center w_i . The competition is incorporated by defining the prior probabilities of the submodels $P(\phi_i)$ as a weighted sum of the input data, passed through a soft-maximum normalization. These prior probabilities give the mixing proportions, π_i . 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 1.2.4). For example, the update rule for maximum likelihood competitive learning [142] 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 [52] 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 feed-back 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 [94] 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” [96]. 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.

$$C(\alpha, d) = C(\alpha)C(d|\alpha) \quad (1.4)$$

The algorithm minimizes expected cost over all of the hidden states

$$E(C(\alpha, d)) = \sum_{\alpha} Q(\alpha|d)C(\alpha, d) \quad (1.5)$$

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 many 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 [94]. It can be shown that Bayesian inference models are equivalent to a minimum description length principle [139]. 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 be 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.

1.2.2 Redundancy reduction as an organizational principle

Redundancy reduction has been proposed as a general organizational principle for unsupervised learning. Horace Barlow [8] 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 1.1 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" [9].

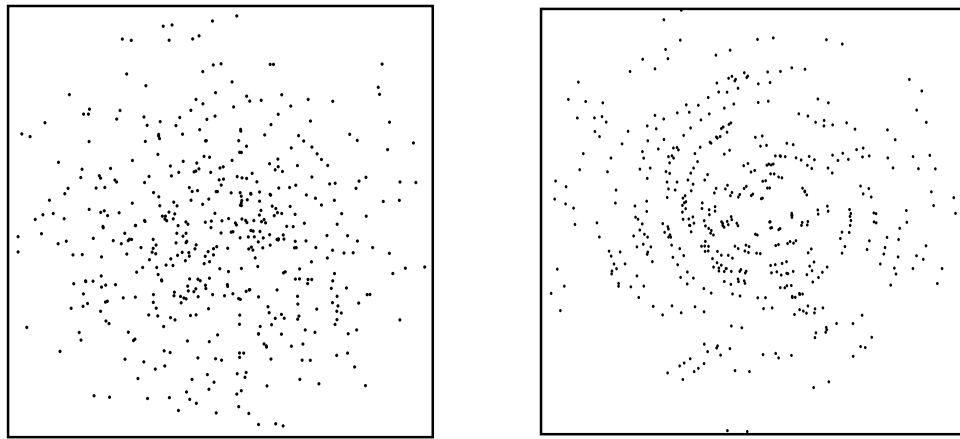


Figure 1.1: The percept of structure is driven by the dependencies. LEFT: A set of points selected from a Gaussian distribution. RIGHT: Half of the points were selected from a Gaussian distribution, and the other half were generated by rotating the points 5° about the centroid of the distribution. Figure inspired by Barlow (1989).

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. Learning such a transformation is equivalent to modeling the prior knowledge of the statistical dependencies in the input [8]. 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. 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 [8]. Contrast gain control, which has been described in V1 [91], takes account of changes in the variance of the input signals.

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 [179]:

$$I(x) = -\log_2 P(x) \quad (1.6)$$

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:

$$H(x) = -\sum P(x) \log_2 P(x) \quad (1.7)$$

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

$$H(x_1, x_2) = H(x_1) + H(x_2) - I(x_1, x_2) \quad (1.8)$$

where $I(x_1, x_2)$ is the mutual information between x_1 and x_2 , which is calculated from 1.6 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 (1.8). 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 [70]. 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 [7] 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 [6] describes evolutionary advantages of efficient coding such as coping with information bottlenecks due to limited bandwidth and limited dynamic range. Atick argues 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.

The large monopolar cells (LMC) in the blowfly compound eye eliminate inefficiency due to unequal use of neural response levels [116]. The most efficient response gain is the one such that the probability distribution of the outputs is constant for all output states. The solution is to match the gain of the transfer function to the cumulative probability density of the input. Laughlin [116] 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 $1/f$ [69]. 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 cpd. 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 [127] 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 the signal-to-noise. 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 $\sigma \left(P^{\frac{1}{3}}(x) \right)$. Macleod and von der Twer found that the pleistochrome based on the distribution of cone responses along the $S - (L + M)$ axis accounted well for the spectral sensitivity of the blue-yellow opponent channel.

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 [7] 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.

1.2.3 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 obtained in the rows of A by

$$A = X^T P \quad (1.9)$$

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, $PP^T = Identity$, and right multiplication of 1.9 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 density estimation models described in Section 1.2.1. 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.

1.2.4 Hebbian learning

Hebbian learning is an unsupervised learning rule that was proposed as a model for activity dependent modification of synaptic strengths between neurons [89]. 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 [22] 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 [124]. 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 [145].

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

$$\Delta w_i = \alpha x_i y \quad (1.10)$$

with learning rate α will move the vector $w = [w_1, \dots, w_n]$ 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

$$\Delta w = \alpha y(x - wy) \quad (1.11)$$

This learning rule converges to the unit length principal component. The $-wy^2$ term tends to decrease 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 [146, 177]. 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

$$\Delta w_j = \alpha y_j \left(x - \sum_{k=1}^{j-1} w_k y_k \right) \quad (1.12)$$

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 is updated.

Linsker [124] 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

$$R = I(x, y) = H(y) - H(y|x) \quad (1.13)$$

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, 1.13 is maximized by maximizing the variance of the output [124]. 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 1.13, 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 1.2.7, in the next section, Linsker [123] demonstrated how structured receptive fields 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, implemented in a local learning rule that was subject to constraints. Information maximization has recently been generalized to the multi-unit case [24]. This dissertation examines representations for face images based on information maximization.

1.2.5 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 [22]. 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 [155] 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 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 [17] 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 [20, 21] 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

$$I = 0.5 \log \frac{V(y_a + y_b)}{V(y_a - y_b)} \quad (1.14)$$

where V is the variance function over the training cases. Maximizing I minimized the squared difference between the module outputs relative to how much both modules varied as the input varied. 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 1.14). 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 [20], and to model the locations of discontinuities in depth [18].

Schraudolph and Sejnowski [178] 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 [211] 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 [22].

1.2.6 High-order statistical dependencies

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 and curvature.

Second-order statistics capture the amplitude spectrum of images but not the phase [70]. 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). The remaining information that is not captured by the autocorrelation function, the high order statistics, corresponds to the phase spectrum.¹

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 [149, 162]. 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.

Relation of sparse coding to independence.

Atick argued for compact, decorrelated codes such as PCA because of efficiency of coding. Field [70] 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. Maximizing sparseness of a response distribution is therefore equivalent 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 [70] are also arguments in favor of Barlow's minimum entropy codes [8]. 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

¹Given a translation invariant input, it is not possible to compute any statistics of the phase from the amplitude spectrum (Dan Ruderman, personal communication.)

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 [154, 16].

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 [148, 147] 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.

Independent component analysis

Independent component analysis (ICA) [43] 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 1.2. 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 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 [70].²

Bell and Sejnowski [24] recently developed an algorithm for separating the statistically independent components of a dataset through unsupervised learning. The algorithm is based on the principle of maximum information transfer between sigmoidal neurons. This algorithm generalizes Linsker's information maximization principle [124] 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.

The Bell and Sejnowski algorithm finds a weight matrix W by performing gradient ascent on the joint entropy of the output Y of a nonlinear transfer function g of the input data,

²Thanks to Michael Gray for this observation.

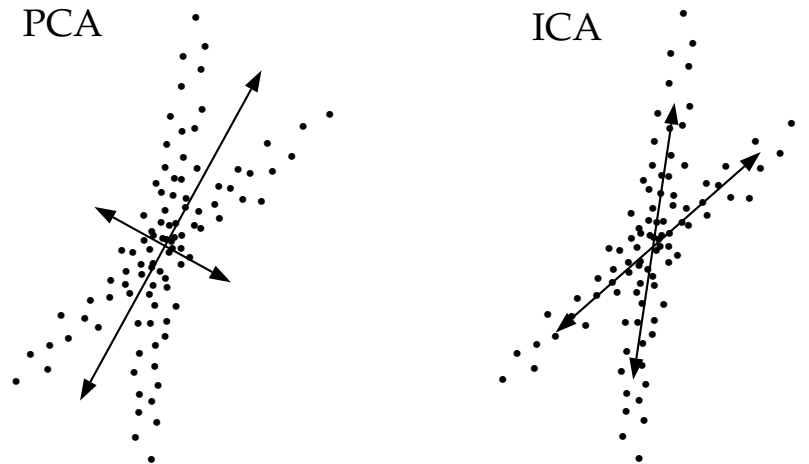


Figure 1.2: Example 2-D data distribution and the corresponding principal component and independent component axes. Figure inspired by Lewicki & Sejnowski (submitted).

$Y = g(WX)$. Maximizing the joint entropy of the output is equivalent to maximizing the mutual information between the input and the output (i.e. maximizing information transfer). This is because $I(X, Y) = H(X) + H(Y) - H(Y|X)$, where only $H(Y)$ depends on the weight matrix W since $H(Y|X)$ is noise. Maximizing the joint entropy of the output encourages the mutual information between the individual outputs to be small (see Equation 1.8). 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 [141, 25].

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 1.8. 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.

Bell & Sejnowski examined the image filters that give independent outputs from natural scenes [25]. 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. Decorrelation mechanisms such as principal components resulted in spatially opponent receptive fields, some of which were oriented, but were not spatially local.

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

[24]. The algorithm has recently been applied to separating the sources of EEG signals [128], and fMRI images [131].

1.2.7 Self-organization of the visual system through correlation sensitive mechanisms

Biological evidence

The gross organization of the visual system appears to be governed by molecular specificity mechanisms during embryogenesis [86]. 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 may be mediated by molecular specificity mechanisms. During postnatal development, the architecture of the visual system continues to become defined, organizing into ocular dominance and orientation columns. The statistical properties of early visual experience and endogenous activity appear to be responsible for shaping the cortical receptive field architecture. (See [189] for a review.)

The NMDA receptor could be the “correlation detector” for Hebbian learning. It opens calcium channels in the post synaptic cell in a manner that 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 [169].

Visual development appears to be closely associated with NMDA gating [44]. 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 [40]. 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 [73].

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 [100]. Colin Blakemore [30] 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 [191]. Blocking post-synaptic activity during monocular deprivation nulls the usual shift in ocular dominance [183, 82]. Stryker demonstrated that ocular dominance segregation depends on asynchronous activity in the two eyes [189]. 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 [53]. Some of organization of ocular dominance and orientation prefer-

ence does occur prenatally. Endogenous activity can account for the segregation of ocular dominance in the lateral geniculate nucleus [5], and endogenous activity tends to be correlated in neighboring retinal ganglion cells [130].

Intrinsic horizontal axon collaterals in the striate cortex of adult cats specifically link columns having the same preferred orientation. Calloway and Katz [37] 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.

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 [123] demonstrated that orientation columns can arise from random input activity in a layered system with 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 Mexican hat shaped.

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 [143]. In Linsker's model, a linear progression of orientation preference would require an isotropic autocorrelation function.

Miller, Keller, and Stryker [133] 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, 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 monocularly 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. A small within-eye correlation was required if there was to be any binocularly at all. Anticorrelation within an eye decreases monocularly, whereas anticorrelation between eyes, such as occurs in conditions of strabismus and monocular deprivation, increases monocularly. 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. Their simulation made a number of experimental predictions. 1. The dependence of ocular dominance on excitatory intracortical connections suggests that muscimol, a GABA agonist, should eliminate ocular dominance organization. 2. Blocking intracortical inhibition with bicuculline should increase patch width up to the size of the thalamo-cortical arbors. 3. Inducing broader within-eye correlations with electrode stimulation should increase monocularly.

Berns, Dayan, and Sejnowski [26] 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 dis-

parity, whereas the more monocular cells that tended to have nonzero disparity.

Obermayer, Blasdel, and Schulten [144] 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, such as 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, (x,y) spatial position, orientation preference, orientation specificity, and ocular dominance. The geometrical patterns arise out of the associated probabilities of the input patterns. 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.

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 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.

1.2.8 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 viewpoint invariant representations.

Földiák [72] 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 1.2.3, 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 Földiák [72] used temporal association to model the development of translation independent orientation detectors such as the complex cells of V1. His model was a two-layer network in which the input layer consisted of sets of local position dependent orientation detectors. This layer was fully connected to four output units. Földiák 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 [130, 132]. 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.

This mechanism can learn viewpoint-tolerant representations when different views of an object are presented in temporal continuity [72, 205, 167, 150, 204]. Földiák 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 [171].

O'Reilly and Johnson [150] 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. O'Reilly and Johnson found that as an object is continuously presented to the network, more and more units are recruited to represent that object. Only unrecruited units and units without saturated connections could respond to the new objects.

Becker [18] showed that the IMAX learning procedure [20], was also able to learn depth from random dot stereograms by applying a *temporal* coherence assumption. 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 [186] 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 straightforward implementation (Stone, personal communication). The two algorithms make the assumption that properties of the visual world such as depth vary slowly in time. Stone [186] 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 [205] applied the assumption of temporal persistence of objects to learn object representations that were invariant to rotations in depth. They 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 as evidence for mental rotation of an internal 3-D model [181].

Weinshall and Edelman modeled the development of viewpoint invariance using idealized objects consisting of paper-clip style objects with labeled vertex locations. The temporal coherence assumption has more recently been applied to learning viewpoint invariant representations of objects in graylevel images [12, 13, 204, 19]. Földiák's learning scheme can be applied in a multi-layer multi-resolution network to learn transformation invariant letter recognition [203], and face recognition that is invariant to rotations in the plane [204]. Becker [19] extended a competitive mixture-of-Gaussians learning model [142] 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 identity. Chapter 5 of this thesis [13] examines the development of representations of faces that are tolerant to rotations in depth in both a feedforward system based on Földiák's learning mechanism, and in a recurrent system in which lateral interconnections formed an attractor network.

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 Földiák, 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 [199] or nonlinear combinations of stored 2-D views [163, 35]. 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 [29, 33]. Face cells in the primate inferior temporal lobe have

been reported with broad pose tuning on the order of $\pm 40^\circ$ [158, 87]. Perrett and colleagues [158], 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 [167, 170]. A spatio-temporal window for Hebbian learning could also be produced by the release of a chemical signal following activity such as nitric oxide [136]. Reciprocal connections between cortical regions [150] 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 of primates [171]. 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 [135]. Macaques were presented a fixed sequence of 97 fractal patterns for 2 weeks. After training, responses to patterns that had been nearby in the training sequence were correlated, although the patterns were presented in random order during the test. 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. 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 [190].

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 [114]. Recognition of novel views of unfamiliar faces was superior when the faces were presented in continuous motion during learning [161]. Stone [187] 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.

1.3 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 [106, 34]. An advantage of the feature-based approach is that it drastically reduces the number of input dimensions. 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 [32]. 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 [34, 115]. Accurate alignment of the faces is critical to the success of template-based approaches. Aligning the face, however, can be more straightforward than locating individual facial landmarks for feature-based representations.

A form of template matching that has achieved success for face recognition is based on principal component analysis of the image pixels [134, 47, 197]. 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” [47] and “Eigenfaces” [197]. 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 face recognition. Representations based on principal component analysis have been applied successfully to recognizing facial identity [47, 197], facial expressions [48, 14, 153], and to classifying the gender of the face [78].

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 [47]. The networks are trained by back-propagation 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 [201] 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 are not necessarily the ones that provide the most information for recognizing facial identity. O’Toole and colleagues [151] 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 [172]. The face cells in the superior temporal sulcus responded most strongly to face images containing energy between 4 and 32 cycles per image.

Principal component analysis is a form of autoassociative memory [200]. The PCA network reproduces the input in the output with minimum squared error. Kohonen [112] 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 [134]. 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 [200].

In such autoassociative networks, a whole face can be recovered from a partial input, thereby acting as content-addressable memory. Cottrell [46] 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 [152] [85]. Such representations have also demonstrated phenomena such as the “other race effect” [152]. 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 [45]. 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 [23]. 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 [180] [84]. 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 the classes impossible.

Penev and Atick [156] recently 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 [153] found that an “Eigenfeature” representation of face images, based in the principal components of image regions containing an 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 [79] 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 [50]. Representations based on the outputs of these filters at multiple spatial scales, orientations, and spatial locations, have been shown to be useful for recognizing facial identity [113]. Relationships have been demonstrated between Gabor filters and statistical independence. Bell & Sejnowski [25] 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 pairwise independent in the presence of divisive normalization [182].

The elastic matching algorithm [113] represents faces by these banks of Gabor filters, and 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 [212, 160].

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 :

$$H^M(i^I) = \frac{a}{2} \sum_{i,j} D_l(L_{ij}^I, L_{ij}^M) - \sum_i S_v(J_i^I, J_i^M) \quad (1.15)$$

where

$$D_l(L_{ij}^I, L_{ij}^M) = (L_{ij}^I - L_{ij}^M)^2 \quad (1.16)$$

$$S_v(J_i^I, J_i^M) = \frac{J_i^I \cdot J_i^M}{\|J_i^I\| \|J_i^M\|} \quad (1.17)$$

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.

The Gabor wavelets, PCA, and independent component analysis (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 [119]. ICA allows the learning of non-orthogonal bases and allows the data to be modeled with non-Gaussian distributions [43]. As noted in Section 1.2.6, there are relationships between Gabor wavelets and the basis functions obtained with ICA [25]. The Gabor wavelets are not specialized to the particular data ensemble, but would be advantageous when the number of data samples is small.

Chapter 2

Independent Component Representations for Face Recognition

2.1 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 [24]. 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 sessions and changes in expression.

This chapter, in part, is a reprint of the following: Bartlett, M.S., Lades, H.M. & Sejnowski, T.J. (in press). "Independent component representations for face recognition," in Rogowitz, B. & Pappas, T., (Eds.) *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology*, vol. 3299, Human Vision and Electronic Imaging III.

2.2 Introduction

Horace Barlow has argued that redundancy provides knowledge [8]. 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 therefore constitute a “suspicious coincidence” in Barlow’s formulation [9]. 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 [8]. Contrast gain control, which has been observed in V1 [91], takes account of changes in the variance of the input signals. 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 [7] 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” [197], “Holons” [48], and “Local Feature Analysis” [156] 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 phase¹ [70, 25]. The Fourier transform of the autocorrelation function of a signal is equal to its power spectrum (square of the amplitude spectrum). The remaining information that is not captured by the autocorrelation function, the high order statistics, corresponds to the phase spectrum. Spatial phase contains the structural information in images that drives human recognition much more strongly than the amplitude spectrum [149, 162]. 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 [43] is a generalization of principal component analy-

¹Given a translation invariant input, it is not possible to compute any statistics of the phase from the amplitude spectrum (Dan Ruderman, personal communication.)

sis that separates the high-order dependencies in the input, in addition to the second-order dependencies. Bell and Sejnowski [24] recently developed an algorithm for separating the statistically independent components of a dataset through unsupervised learning. This algorithm has proven successful for separating randomly mixed auditory signals (the cocktail party problem), and has recently been applied to separating EEG signals [128], fMRI signals [131], and finding image filters that give independent outputs from natural scenes [25].

This paper 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), and found a set of statistically independent source images for a set of face images. These source images comprised a set of independent basis images for the faces, and can be viewed as set of statistically independent facial feature images, in which the pixel values in one feature image cannot be predicted from the pixel values of the other feature 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 [24] and to examine the independent sources of EEG [128] and fMRI data [131]. Under this architecture, the basis images were independent, but the coding variables that represented each face image were not. The second architecture separated pixels across images, and corresponded to the architecture used to find image filters that produced statistically independent outputs from natural scenes [25]. This architecture defined a set of statistically independent coding variables for representing the face images. In other words, we used ICA to find a factorial face code.

Face recognition performance was tested using the FERET database [160]. Face recognition performances using the ICA representations were benchmarked by comparing them to performances using principal component analysis, which is equivalent to the “Eigenface” representation [197, 157].

2.3 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 [116, 24]. Consider the case of a single input, x , and output, y , passed through a nonlinear squashing function, g .

$$u = wx + w_0 \quad y = g(u) = \frac{1}{1 + e^{-u}} \quad (2.1)$$

As illustrated in Figure 2.1, 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 . 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

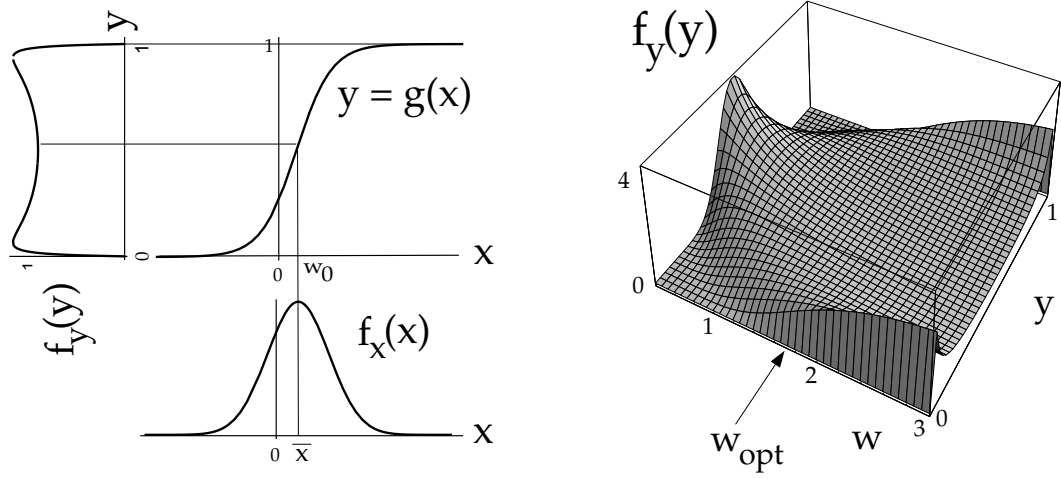


Figure 2.1: Optimal information flow in sigmoidal neurons. The input x is passed through a non-linear function, $g(x)$. The information in the output density $f_y(y)$ depends on matching the mean and variance of $f_x(x)$ to the slope and threshold of $g(x)$. Right: $f_y(y)$ is plotted for different values of the weight, w . The optimal weight, w_{opt} transmits the most information. Figure from Bell & Sejnowski (1995), reprinted with permission from *Neural Computation*, copyright 1995, MIT Press.

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 and the output Y also minimizes the mutual information between the u_i [141, 25]. 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 [24]. For mixtures of super-Gaussian signals, the logistic transfer function has been found to be sufficient to separate the signals [24].

The update rule for the weight matrix, W , for multiple inputs and outputs is given by

$$\Delta W = (I + y' u^T) W \quad (2.2)$$

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

We employed the logistic transfer function, $g(u) = \frac{1}{1+e^{-u}}$, giving $y' = (1 - 2y_i)$. The algorithm includes a “sphering” step prior to learning [25]. 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:

$$W_z = 2 * \langle X X^T \rangle^{-\frac{1}{2}}. \quad (2.3)$$

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 [25].

The difference between ICA and PCA is illustrated as follows. Consider a set of data points derived from two underlying distributions as shown in Figure 2.2. Principal component analysis encodes second order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. PCA models the data as a multivariate Gaussian and would place an orthogonal set of axes such that 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 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 [70].

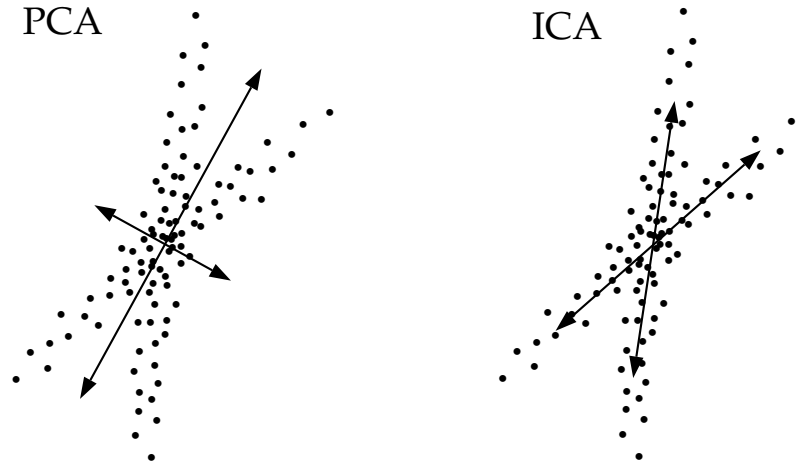


Figure 2.2: Example 2-D data distribution and corresponding principal component and independent component axes. Figure inspired by Lewicki & Sejnowski (submitted).

2.4 Independent Component Representations of Face Images

2.4.1 Statistically independent basis images

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 2.3. 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 .

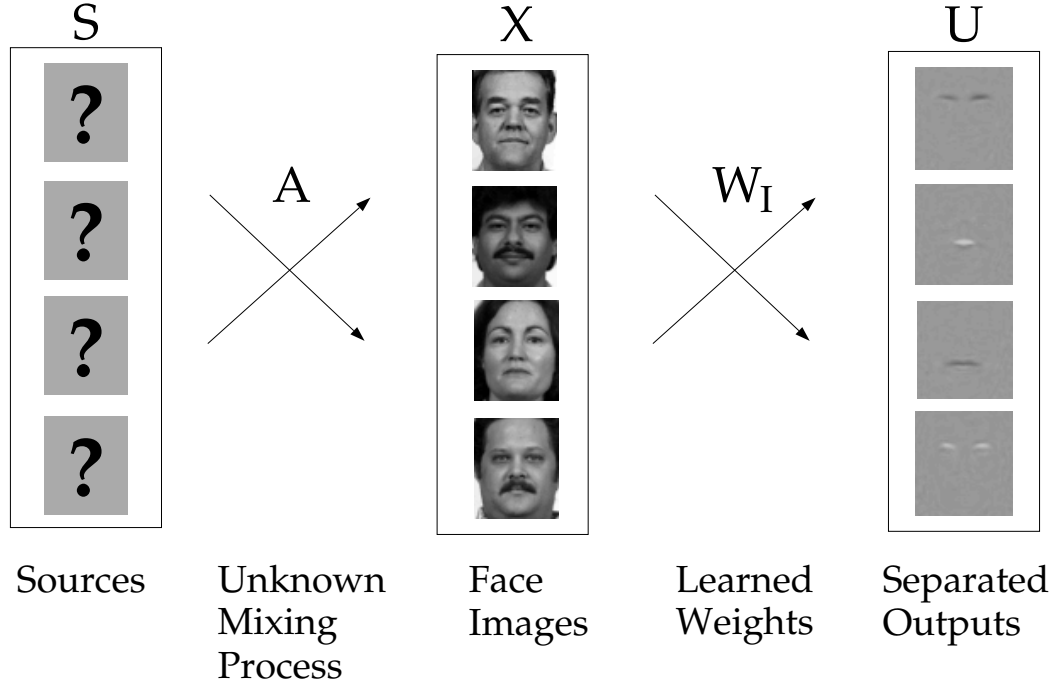


Figure 2.3: Image synthesis model. For finding a set of independent component images, the images in X are considered to be a linear combination of statistically independent basis images, S , where A is an unknown mixing matrix. The basis images were recovered by a matrix of learned filters, W_I , that produced statistically independent outputs, U .

This synthesis model is related to the one used to perform blind separation on an unknown mixture of auditory signals [24] and to separate the sources of EEG signals [128] and fMRI images [131].

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 2.4). 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 comprised each face image, as shown in Figure 2.4. The matrix of coefficients, B , was obtained from the mixing matrix $A \triangleq W_I^{-1}$.

The number of independent components found by the ICA algorithm corresponds to the dimensionality of the input. 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 sta-

$$\text{Face Image} = b_1 * u_1 + b_2 * u_2 + \dots + b_n * u_n$$

$$\text{ICA representation} = (b_1, b_2, \dots, b_n)$$

Figure 2.4: 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} .

tistically 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 [131], 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 . 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

$$W_I * P_m^T = U \quad \Rightarrow \quad P_m^T = W_I^{-1} U. \quad (2.4)$$

Therefore

$$X_{rec} = R_m * P_m^T \quad \Rightarrow \quad X_{rec} = R_m * W_I^{-1} U. \quad (2.5)$$

where W_Z was the sphering matrix defined in Equation 2.3. 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

$$B = R_m * W_I^{-1}. \quad (2.6)$$

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 $B_{test} = R_{test} * W_I^{-1}$.

2.4.2 Independence in face space versus pixel space

The analysis in Section 2.4.1 produced statistically independent basis images. The ICA algorithm separated images across pixel location (see Figure 2.5 Top Left.) Each pixel location was an observation which took on different grayvalues for each of the faces. This is illustrated in Figure 2.5 (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 space of the face images, and each face was represented by the coefficients for the linear combination of these independent template images that comprised each face image.

The correspondence of the ICA-basis representation (ICA1) with the principal component representation is that the Eigenface images are pixelwise uncorrelated. PCA represents faces as a linear combination of uncorrelated template images, whereas ICA represents faces as a linear combination of independent template images. The Eigenfaces are uncorrelated because 1. They are orthogonal by definition, and 2. The Eigenfaces are themselves a set of PCA coefficients obtained by considering each pixel location an observation and each face image a measure. Let v_i denote the eigenvectors of $X^T X$. The eigenvectors u_i of $X X^T$ can be obtained by $X v_i$ [197].

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 2.5 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.

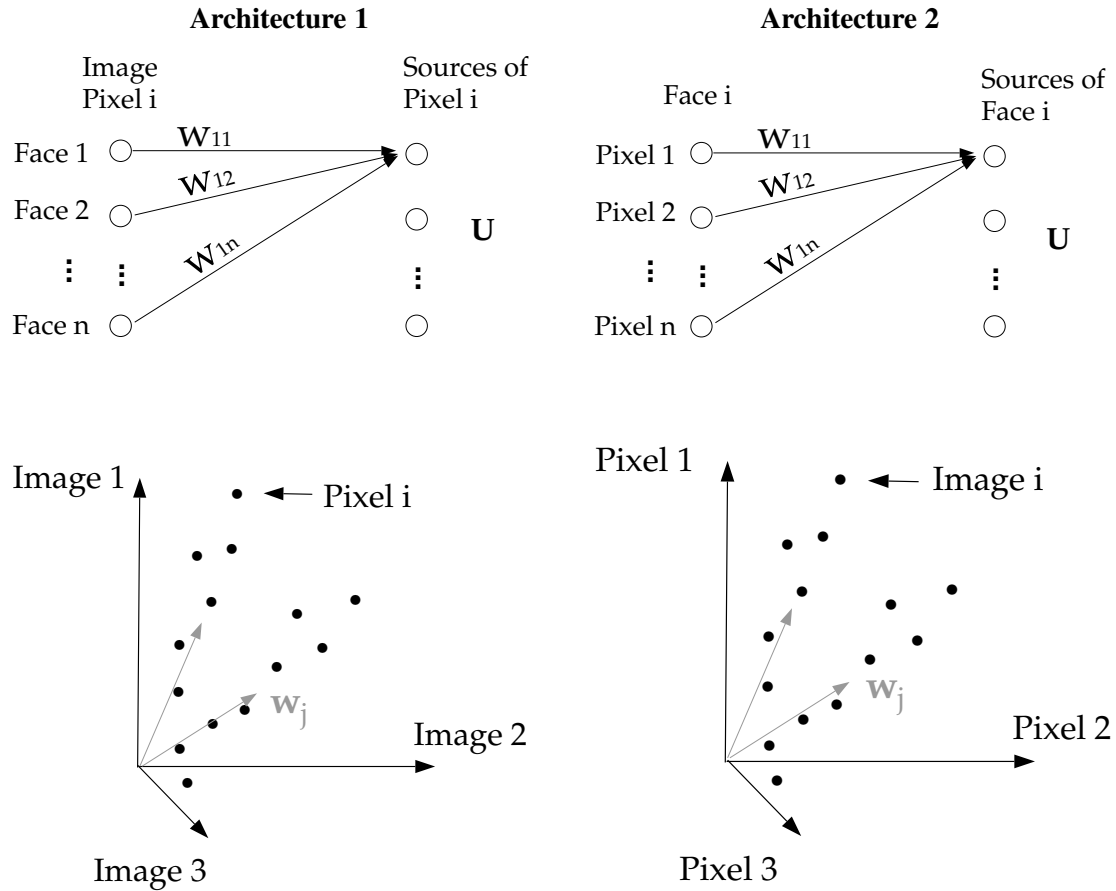


Figure 2.5: Two architectures for performing ICA on images. LEFT: Architecture for finding statistically independent basis images. Top Left: Performing source separation on the face images produced independent component images in the rows of U . Bottom left: The grayvalues at pixel location i are plotted 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.

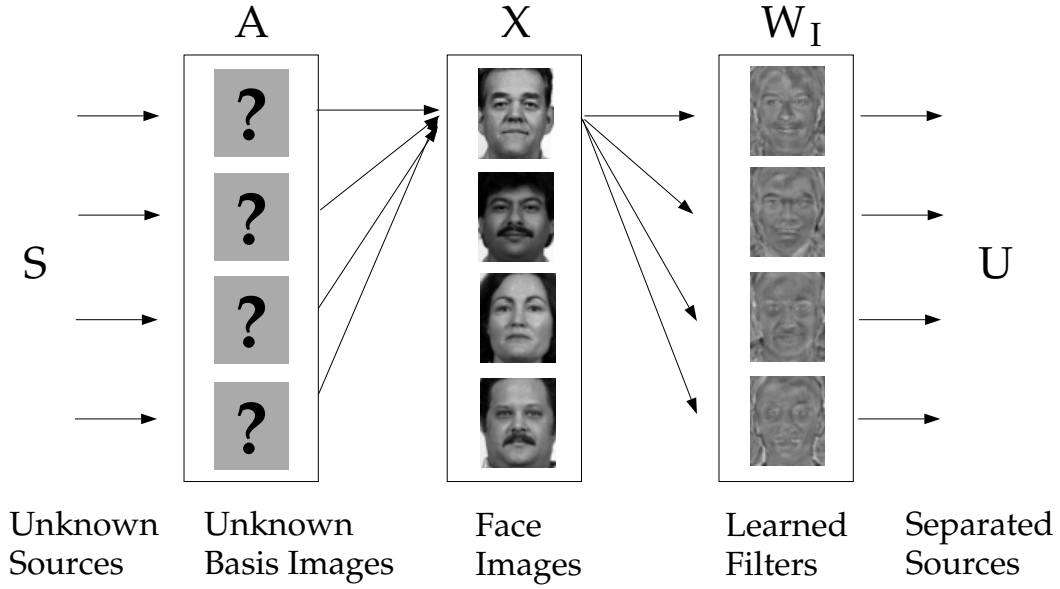


Figure 2.6: Image synthesis model for Architecture 2, based on Olshausen & Field (1996) and Bell & Sejnowski (1997). Each image in the dataset was considered to be a linear combination of underlying basis images in the matrix A . The basis images were each associated with a set of independent “causes”, 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 .

2.4.3 A factorial face code

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 2.5 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 2.7. This model is similar to the the model in Figure 2.3, except that we now 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 2.3 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 [148], and was also employed by Bell and Sejnowski [25] 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

$$x = u_1 * a_1 + u_2 * a_2 + \dots + u_n * a_n$$

$$\text{ICA factorial representation} = (u_1, u_2, \dots, u_n)$$

Figure 2.7: The factorial code representation consisted of the independent coefficients, u , for the linear combination of basis images in A that comprised each face image x .

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 2.7). 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.

2.5 Face Recognition Performance

Face recognition performance was evaluated for the two ICA representations using the FERET face database [160]. 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 2.8. The two 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 session, same expression; and different session, different expression (see Table 2.1).



Figure 2.8: 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.

Image Set	Condition		Number of Images
Training Set	Session I	50% neutral 50% other	425
Test Set 1	Same Session	Different Expression	421
Test Set 2	Different Session	Same Expression	45
Test Set 3	Different Session	Different Expression	43

Table 2.1: 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.

2.5.1 Independent basis architecture

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. 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 2.3, and the weights, W , were updated according to Equation 2.2 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 2.9 shows a subset of 25 source images. A set of principal component basis images (PCA axes), are shown in Figure 2.10 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 [25].

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 2.6, and coefficients for the test images were contained in the rows of $B_{test} = R_{Test} * W_I^{-1}$ where $R_{Test} = X_{test_i} * P_{200}$.

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:

$$d = \frac{\mathbf{b}_{test} \cdot \mathbf{b}_{train}}{\|\mathbf{b}_{test}\| \|\mathbf{b}_{train}\|}. \quad (2.7)$$

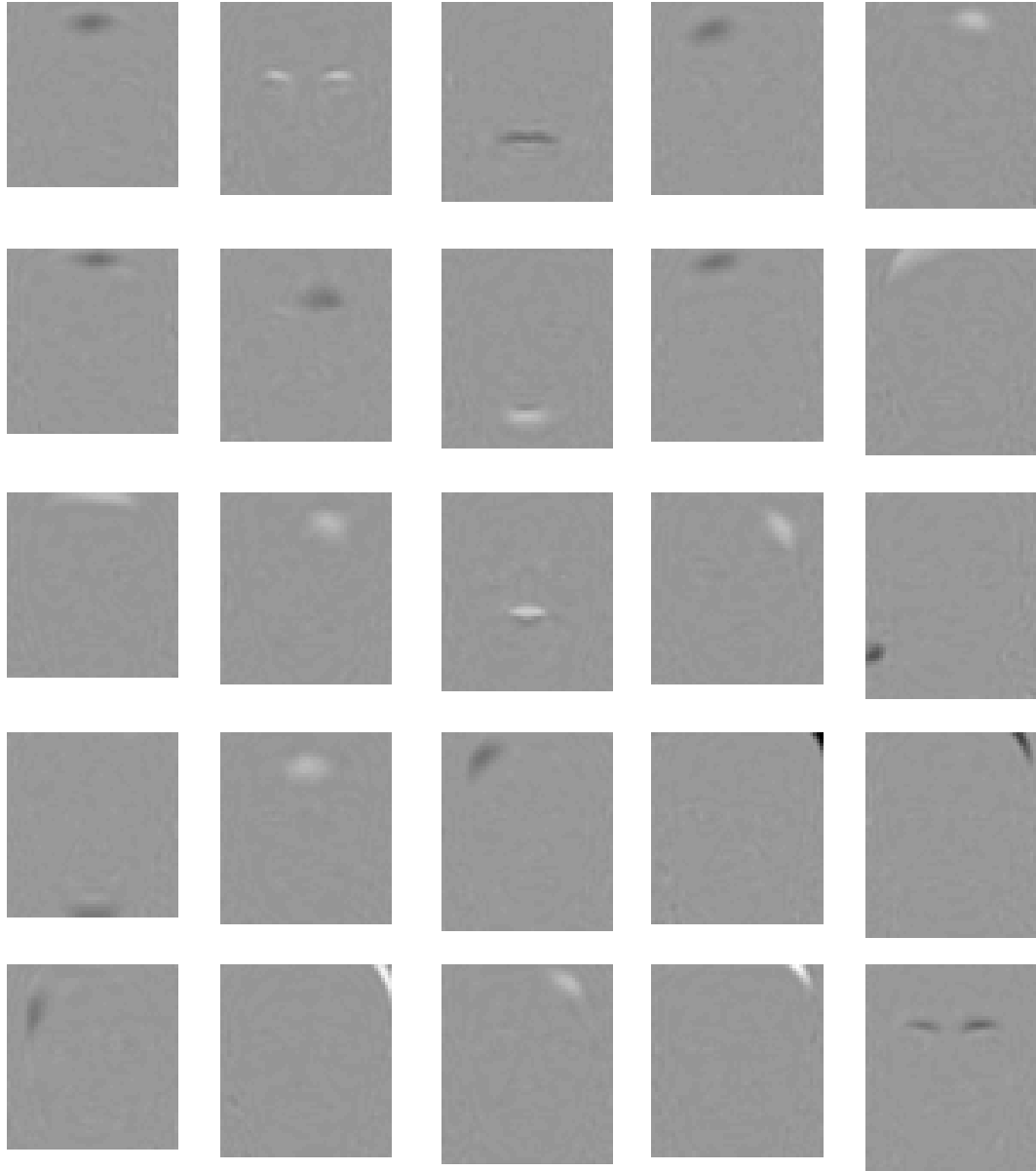


Figure 2.9: Twenty-five independent components of the image set obtained by Architecture 1, which provide a set of statistically independent basis images (rows of U in Figure 2.3). Independent components are ordered by the class discriminability ratio, r (Equation 2.8).



Figure 2.10: 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 performance for the principal component representation was evaluated by an identical procedure, using the principal component coefficients contained in the rows of R_{200} . Figure 2.11 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 a recent Eigenface representation used by Pentland, Moghaddam and Starner [157]. 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.

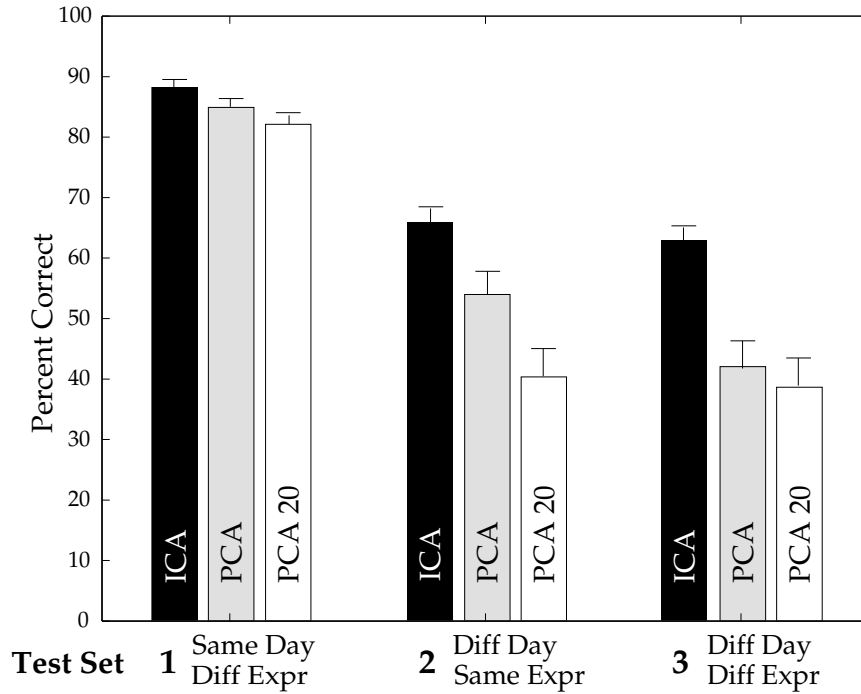


Figure 2.11: Percent correct face recognition for the ICA representation using 200 independent components, the PCA representation using 200 principal components, and the PCA representation using 20 principal components. Groups are performances for Test Set 1, Test Set 2, and Test Set 3. Error bars are one standard deviation of the estimate of the success rate for a Bernoulli distribution.

Face recognition performances for the PCA and ICA representations were next compared by selecting subsets of the 200 components by class discriminability. Let \bar{x} be the overall mean of a coefficient b_k across all faces, and \bar{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:

$$r = \frac{\sigma_{between}}{\sigma_{within}} \quad (2.8)$$

where $\sigma_{between} = \sum_j (\bar{x}_j - \bar{x})^2$ is the variance of the j class means, and $\sigma_{within} = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ 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 images from the corresponding test set from the analysis. Both the PCA and ICA coefficients were then ordered by the magnitude of r . Figure 2.12 (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 2.12 (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. 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.

2.5.2 Factorial code architecture

ICA was next performed on the face images using Architecture 2 to find independent coding variables across images. 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 200×200 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 2.13, 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 2.9). 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.

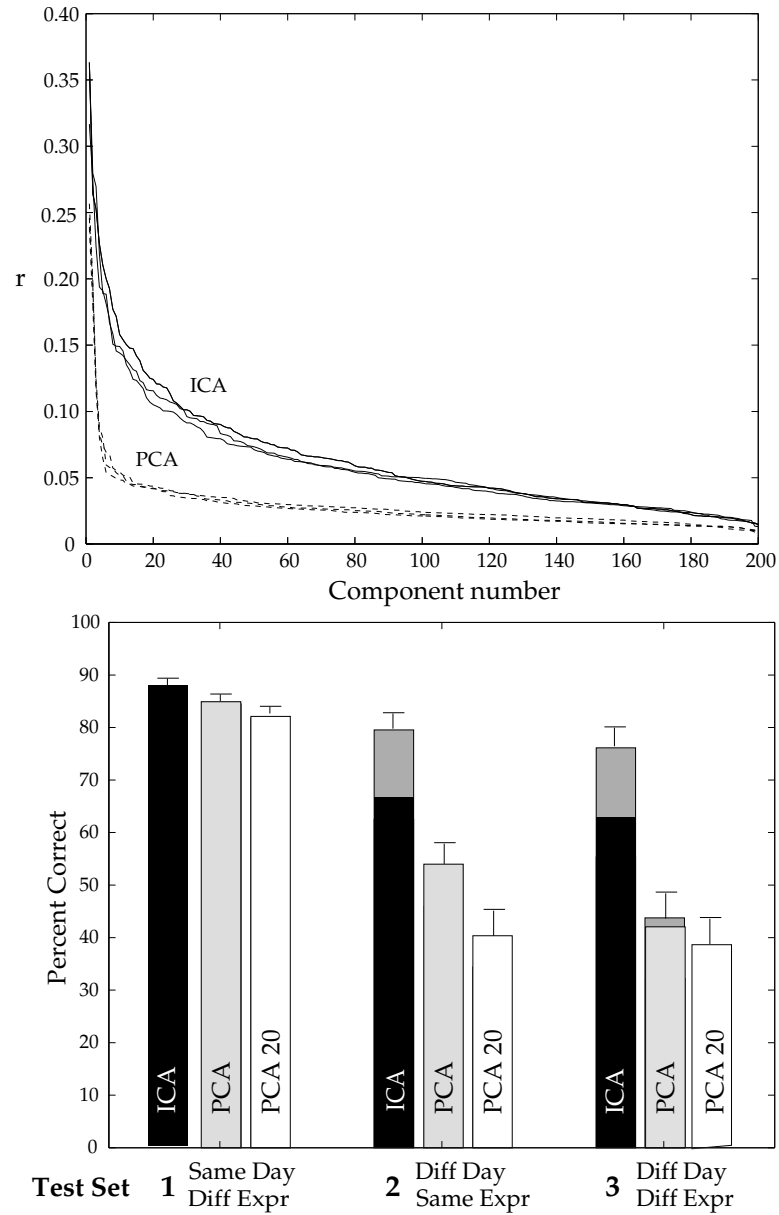


Figure 2.12: Selection of components by class discriminability. Top: Discriminability of the ICA coefficients (solid lines) and discriminability of the PCA components (dotted lines) for the three test cases. Components were sorted by the magnitude of r . Bottom: Improvement in face recognition performance 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.



Figure 2.13: Basis images for the ICA factorial representation (columns of $A = W_I^{-1}$) obtained with Architecture 2. (See Figure 2.6).

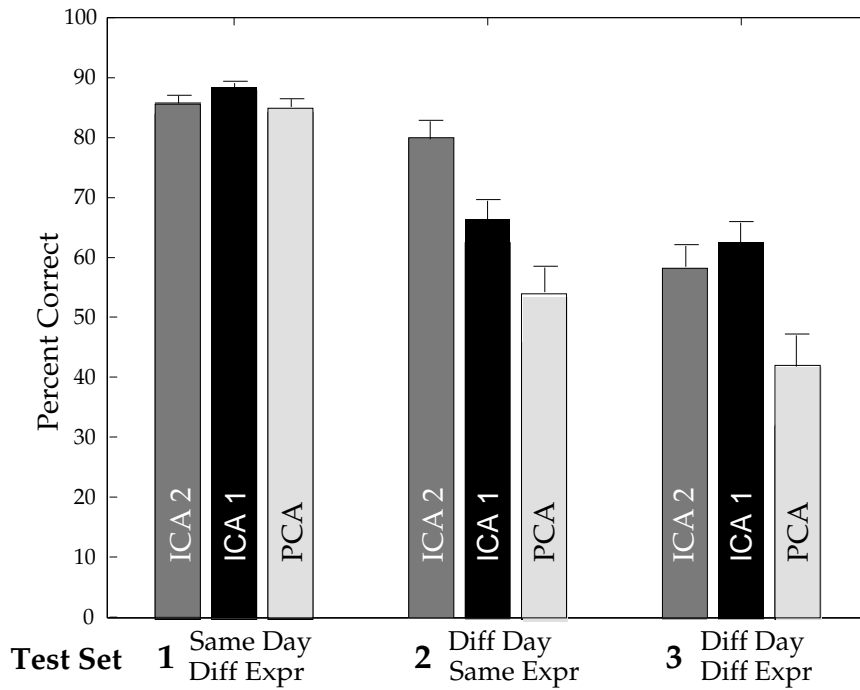


Figure 2.14: Recognition performance of the factorial code ICA representation (ICA2) using all 200 coefficients, compared to the ICA independent basis representation (ICA1), and the PCA representation, also with 200 coefficients.

Face recognition performance was again evaluated by the nearest neighbor procedure. Figure 2.14 compares the face recognition performance using the ICA factorial code representation to the independent basis representation of Section 2.4.1 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 changes in session. 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.

Class discriminability of the 200 ICA factorial coefficients was calculated according to Equation 2.8. 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 2.15). The difference in performance between ICA1 and ICA2 for Test Set 3 following the discriminability analysis just misses significance ($Z = 1.88, p = 0.06$).

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

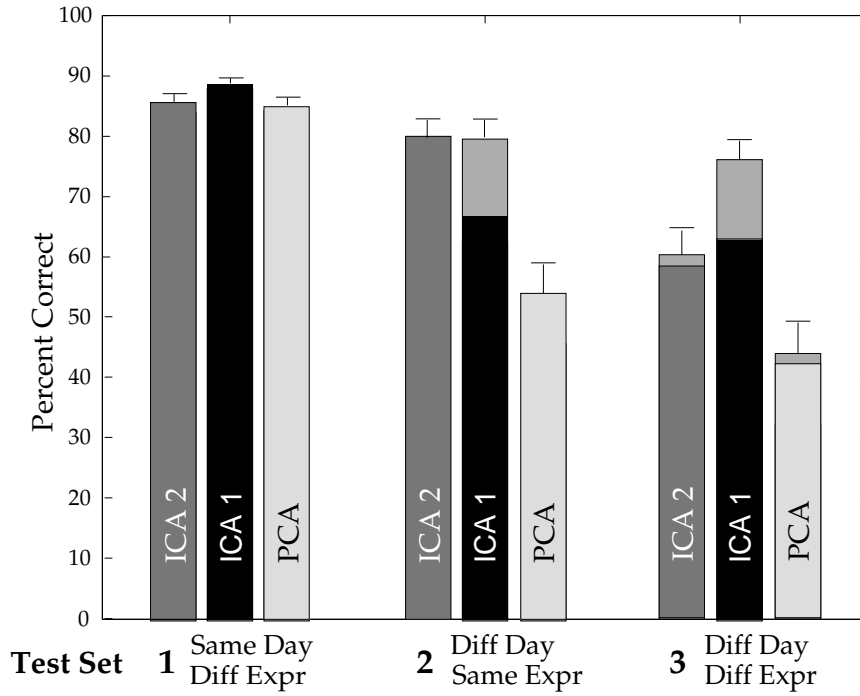


Figure 2.15: Improvement in recognition performance of the two ICA representations and the PCA representation by selecting subsets of components by class discriminability. Gray extensions show improvement over recognition performance using all 200 coefficients.

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.

2.6 Examination of the ICA Representations

2.6.1 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

$$I(u_1, u_2) = \frac{H(u_1) + H(u_2) - H(u_1, u_2)}{H(u_1)} \quad (2.9)$$

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

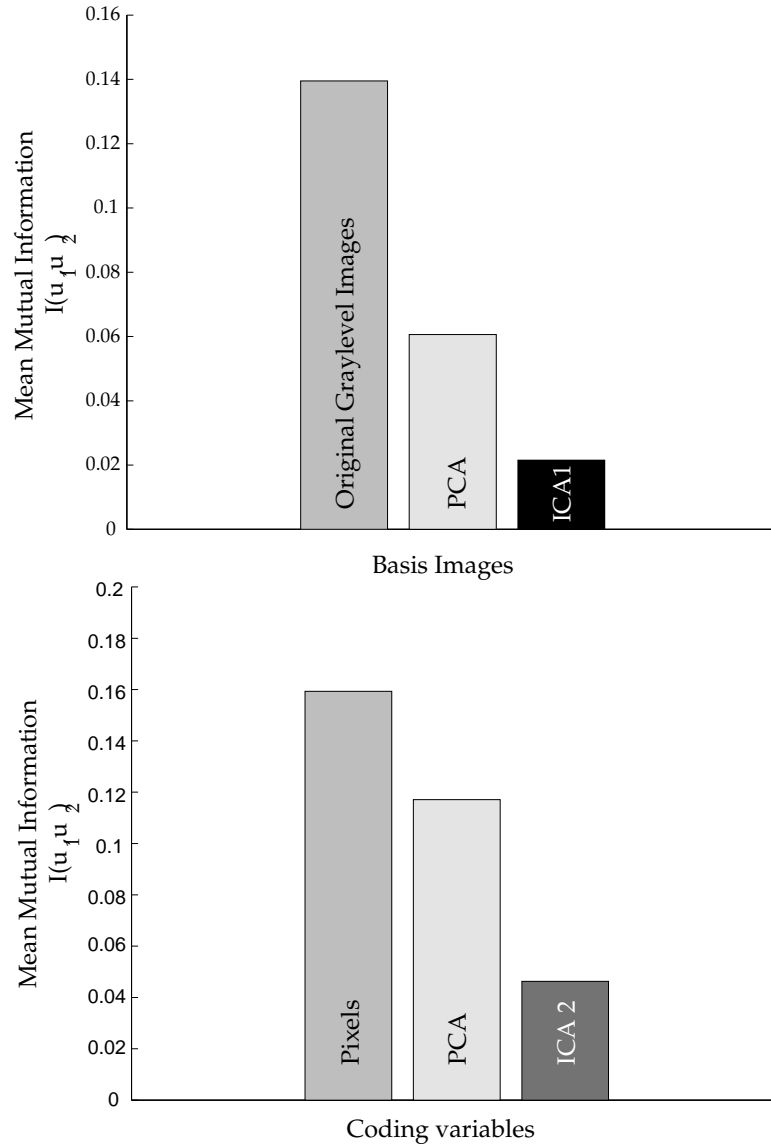


Figure 2.16: Pairwise mutual information. Top: Mean mutual information between basis images. Mutual information was measured between pairs of graylevel images, principal component images, and independent basis images obtained by Architecture 1. Bottom: Mean mutual Information between coding variables. Mutual information was measured between pairs of image pixels in graylevel images, PCA coefficients, and ICA coefficients obtained by Architecture 2.

Figure 2.16 (Top) 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 sources, or the large number of free parameters to be estimated relative to the number of training images.

Figure 2.16 (Bottom) 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.

2.6.2 Sparseness

Field [70] 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 individual outputs. This is equivalent to the minimum entropy codes discussed by Barlow [8]. A transformation that minimizes the entropy of the individual outputs encourages statistical independence between the outputs.²

Given the relationship between sparse codes and minimum entropy, the advantages for sparse codes as outlined by Field in [70] mirror the arguments for independence presented by Barlow in [8]. 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 [154, 16].

²Information 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.

The probability densities for the values of the coefficients of the two ICA representations and the PCA representation are shown in Figure 2.17. 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:

$$kurtosis = \frac{\sum_i (b_i - \bar{b})^4}{\left(\sum_i (b_i - \bar{b})^2\right)^2} - 3 \quad (2.10)$$

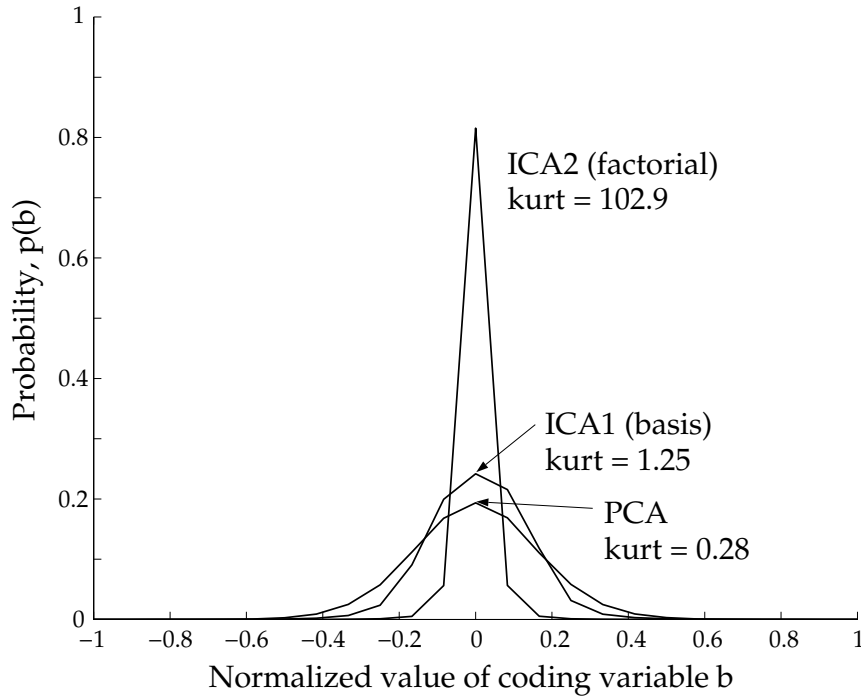


Figure 2.17: Kurtosis (sparseness) of ICA and PCA representations.

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.

2.6.3 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 2.18.



Figure 2.18: 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.

When the two algorithms made errors, however, they did not confuse the same *pairs* of images. 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 gallery image was defined as $d_1 + d_2$, where d_1 and d_2 correspond to the similarity measure d in Equation 2.7 for ICA1 and ICA2. 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 2.19. 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 ($Z = 2.7, p < 0.01$; $Z = 3.7, p < .001$; $Z = 3.7, p < .001$).

2.7 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 “Eigen-faces” and “Holons” are based on principal component analysis, which separates the second-order

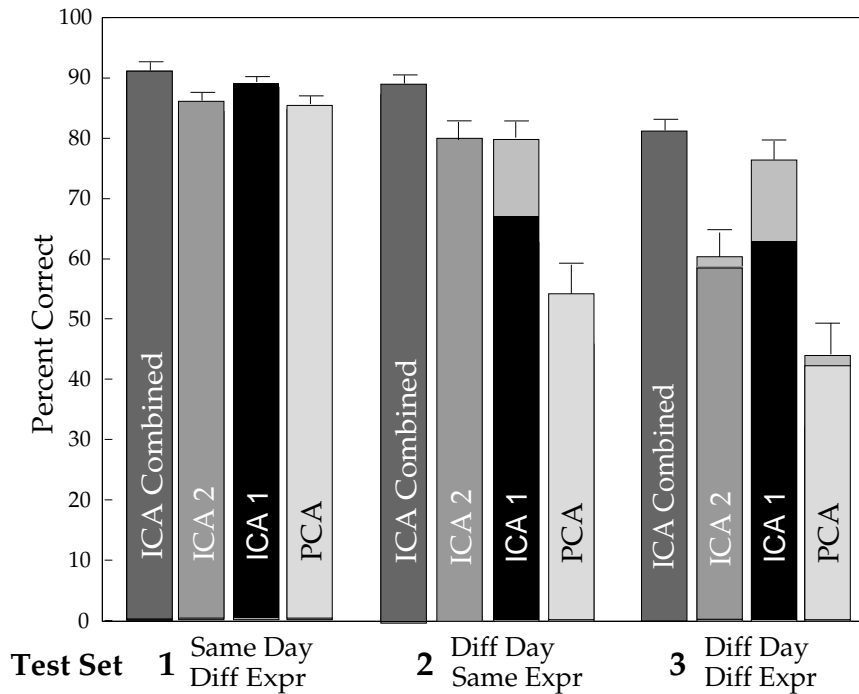


Figure 2.19: Face recognition performance of the combined ICA classifier, compared to the individual classifiers for ICA1 and ICA2, and PCA.

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 (ICA) 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. The ICA representations embodied a prior that these image 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 statistical 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 [197], 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 representations 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 [157]. 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.

The ICA representation of faces presented in Section 2.4.1 is related to the method of local feature analysis (LFA), which is a topographic representation based on principal components analysis. The LFA kernels are found by performing zero phase whitening (Equation 2.3) on the principal component axes, followed by a rotation to topographic correspondence with pixel location. Convolution of an image with these kernels is equivalent to reconstructing the image with whitened principal component coefficients. In LFA, the kernels are matched to the second-order statistics of the input ensemble, whereas in the ICA representation, the kernels are matched to the high-order statistics of the ensemble as well as the second-order statistics. Both methods produce local filters from an objective of redundancy reduction.

In Section 2.4.3, 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 [8, 6]. According to the arguments of both Field [70] and Barlow [8], 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 2.16 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 neighbors 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.

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 and [25]. 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 [122].

The principle of independence, if not the specific learning algorithm employed here [25], may have relevance to face and object representations in the brain. Horace Barlow [8] and

Joseph Atick [6] have argued for redundancy reduction as a general coding strategy in the brain. This notion is supported by the findings of Bell and Sejnowski [25] 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 [148, 147] obtained a similar result with a sparseness objective, where there is a close information theoretic relationship between sparseness and independence [8, 25]. It has also been shown that Gabor filter outputs of natural images are pairwise independent in the presence of divisive normalization such as the contrast gain control mechanisms proposed for V1 simple cells [182, 91]. The finding in this paper that face representations derived from independent component analysis give superior face recognition performance to representations that remove only the second-order redundancies supports arguments that independence is a good strategy for high-level object recognition.

Acknowledgments

We are 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.

This chapter, in part, is a reprint of material that will appear in *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, Vol. 3299, B. Rogowitz & T. Pappas, (Eds.), SPIE, in press. The dissertation author was the primary investigator of this paper. Figure 1.1 was reprinted with permission from *Neural Computation* 7, copyright 1995, MIT Press.

Chapter 3

Measuring Facial Expressions by Computer Image Analysis

3.1 Abstract

Facial expressions provide an important behavioral measure for the study of emotion, cognitive processes, and social interaction. The Facial Action Coding System, [62], is an objective method for quantifying facial movement in terms of component actions. We have applied computer image analysis to the problem of automatically detecting facial actions in sequences of images. Three approaches were compared: Holistic spatial analysis, 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. The hybrid system outperformed human non-experts on this task, and performed as well as highly trained experts. An automated system would make facial expression measurement more widely accessible as a research tool in behavioral science and investigations of the neural substrates of emotion.

3.2 Introduction

Facial expression measurement provides an indicator of emotion activity that is less intrusive than EEG, EMG, ANS or brain imaging measurements, and 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 [64, 66, 67]). Recent advances in computer vision and neural networks open up the possibility of automatic measurement of facial signals. An automated system would make facial expression measurement more widely accessible as a research tool in behavioral science and medicine, and would provide alternative measures of visual stimuli and behavioral responses in psychophysiological investigations into the neural substrates of emotion and facial expression recognition.

Copyright 1988, Society for Psychophysiological Research. Reprinted with permission from *Psychophysiology*, in press, 1998.

3.2.1 Measurement of facial signals

Most research on facial expressions has not measured facial behavior itself, but instead has measured the information that observers were able to infer from looking at facial expressions (see [175]). The Facial Action Coding System (FACS) [62] was developed to enable studies which directly measured facial behavior itself. 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 differences between these two approaches to the study of facial expression (observer inference vs. facial measurement) were discussed and the literature reviewed in [56, 57].

FACS allows precise specification of the morphology and the dynamics of facial movement. FACS was developed by determining from palpation, knowledge of anatomy, videotapes, and photographs how the contraction of each of the facial muscles changed the appearance of the face. Ekman and Friesen defined 46 Action Units, or AUs, to correspond to each independent motion of the face. FACS is coded from video, and a trained human FACS coder decomposes an observed expression into the specific AUs that occurred, and their duration, onset, and offset time. More than 300 people worldwide have achieved inter-coder agreement on the Facial Action Coding System.

Electromyography has served as a useful complement to overt facial action coding systems (see [36] for a review). Facial EMG measures facial behavior directly, but it is not comprehensive and unlike the methods considered here, it is highly obtrusive. Izard's Maximally Discriminative Affect Coding System (MAX) [101], is an alternative facial coding system in which the units are formulated in terms of appearances that are relevant to eight specific emotions, rather than in terms of individual muscles. Unlike FACS, MAX does not exhaustively measure all facial motions. The facial actions which MAX specifies comprise a subset of the facial actions in the FACS emotion dictionary.

In recent years a number of studies have appeared showing the rich variety of information that can be obtained by using FACS (see [67] for a review). Examples include evidence of a facial signal for embarrassment [108]; differences between genuine and simulated pain [49]; cross cultural differences in how infants respond to restraint [39]; signs of psychopathology [185]; and differences between the facial signals of suicidal and non-suicidally depressed patients [92]. Dynamic signals contain information for discriminating genuine expressions of emotion from false ones (e.g. [63]).

Although FACS is a promising approach, a major impediment to its widespread use 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. 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 industry. An automated system would not only increase the speed of coding, it would also improve the reliability, precision, and temporal resolution of facial measurement.

3.2.2 Analysis of facial signals by computer

Some success has been achieved for automatic detection of facial actions by tracking the positions of dots attached to the face [93, 105]. A system that detects facial actions from image sequences without requiring application of dots to the subjects face would have much broader utility. Efforts have recently turned to measuring facial actions by image processing of video sequences [14, 42]. This paper explores and compares methods for classifying facial actions in image sequences of faces.¹

Recent advances have been made in computer vision for automatic recognition of facial expressions in images. The approaches that have been explored include analysis of facial motion [129, 206, 173, 68], measurements of the shapes and facial features and their spatial arrangements [115], holistic spatial pattern analysis using techniques based on principal components analysis [48, 153, 115] and methods for relating face images to physical models of the facial skin and musculature [129, 193, 121, 68]. These systems demonstrate approaches to face image analysis that are applicable to the present goals and are reviewed below, but the systems themselves are of limited use for behavioral and psychophysiological research.

Facial action codes versus emotion categories

Most of the computer vision systems for recognizing facial expressions attempt to classify expressions into a few broad categories of emotion, such as happy, sad, or surprised. The evidence for seven universal facial expressions (see [59] for a review), does not imply that these emotion categories are sufficient to describe all facial expressions [83]. If automated facial measurement were to be constructed simply in terms of seven elementary emotional categories, much important information would be lost: blends of two emotions, variations within an emotional category (eg. vengeance vs. resentment), variations in intensity (annoyance vs. fury), conversational signals, and idiosyncratic facial movements.

Systems that only produced emotion category labels also could not be used in investigations of facial behavior itself. Several computer vision systems explicitly parameterize facial movement [206], and relate facial movements to the underlying facial musculature [129, 68], but these descriptions are not readily interpretable in terms of facial action codes. It is unknown whether these descriptions are sufficient for describing the full range of facial behavior, and the relationship between these measures and internal state has not been established. A large body of empirical data already exists demonstrating the relationship of facial action codes to emotions, emotion intensity, variations, blends, and conversational signals.

Analysis of facial motion

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 [129] used optic flow to estimate the activity in 12 of the 44 facial muscles. For each

¹A brief report of this work appeared in [14].

muscle he defined a window in the face image and an axis along which each muscle expands and contracts. The mean similarity of the flow vectors inside the window to this axis provided a coarse estimate of the activity of the muscle. Yacoob & Davis [206] 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 [173] 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 [28] 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. The system most closely related system to this approach is that of Cohn et al. [42], who are building a system to classify facial actions by motion tracking of manually located facial features in the initial image. 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.

Model-based techniques

Several facial expression recognition systems have employed explicit physical models of the face [129, 193, 121, 68]. Essa & Pentland [68] extended a detailed anatomical and physical model of the face developed by Terzopoulos and Waters [193] 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 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 would be difficult to accurately account for all of them in a deterministic model. In this paper we take a neural network approach to image analysis in which facial action classes are learned directly from example image sequences of the actions, bypassing the physical model.

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 [106, 34]. Lanitis, Taylor, & Cootes [115] 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. 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 [32, 33].

Holistic analysis

The 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 [138]. An example of this approach is template matching. Templates capture information about configuration and shape that can be difficult to parameterize. In related neural network approaches to image analysis, the physical properties relevant to the classification need not be specified in advance, and can be learned from the statistics of the image set. This is particularly useful when the specific features relevant to the classification are unknown [200].

One holistic spatial representation is based on the principal components of the image pixels [47, 197]. 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 which have been labeled “Holons” [47] and “Eigen-faces” [197]. 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 components analysis has been applied successfully to recognizing both facial identity [47, 197], and facial expressions [48, 14, 153]. Another holistic spatial representation is obtained by a class-specific linear projection of the image pixels [23]. Accurate alignment of the faces is critical to the success of such image-based approaches. Feature-based and template-based methods need not be mutually exclusive. Lanitis, Taylor, & Cootes, [115], 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.

3.3 Automating the Facial Action Coding System (FACS)

We explored three different methods for classifying facial actions that were suited to detecting different kinds of image cues: holistic spatial analysis based on principal components, 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. One benchmark for the performances of the automated systems was provided by the ability of naive human subjects to classify the same images. A second benchmark was provided by the agreement rates of expert coders on these images.

3.3.1 Methods

Image Database

We collected a database of image sequences of subjects performing specified 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 3.1). The database therefore contained examples of the facial actions at low and medium magnitude as well as at high magnitude.² Trained FACS experts provided demonstrations and instructions to subjects on how to perform each action. 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.

For this investigation, we used data from 20 subjects and attempted to classify the six individual upper face actions illustrated in Figure 3.2. 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 [62]. 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 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 3.2. 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 [138].

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

²The term “magnitude” replaces the term “intensity” used in FACS to avoid confusion with image intensity.



Figure 3.1: 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.

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.

Holistic spatial analysis

We first evaluated the ability of a back-propagation network to classify facial actions given principal components of graylevel images as input. This approach is based on [48] and [197], with the primary distinction in that we performed principal components 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. The principal component axes of the difference image data were then calculated by finding the eigenvectors of the pixelwise covariance matrix. The axes were ordered by the magnitude of the corresponding eigenvalue. Figure 3.3 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 [88]). The activities of the hidden and output units were calculated sequentially as the weighted sum of their inputs, passed through a sigmoidal hyperbolic tangent transfer function. The network was trained by back-propagation of error to output 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.

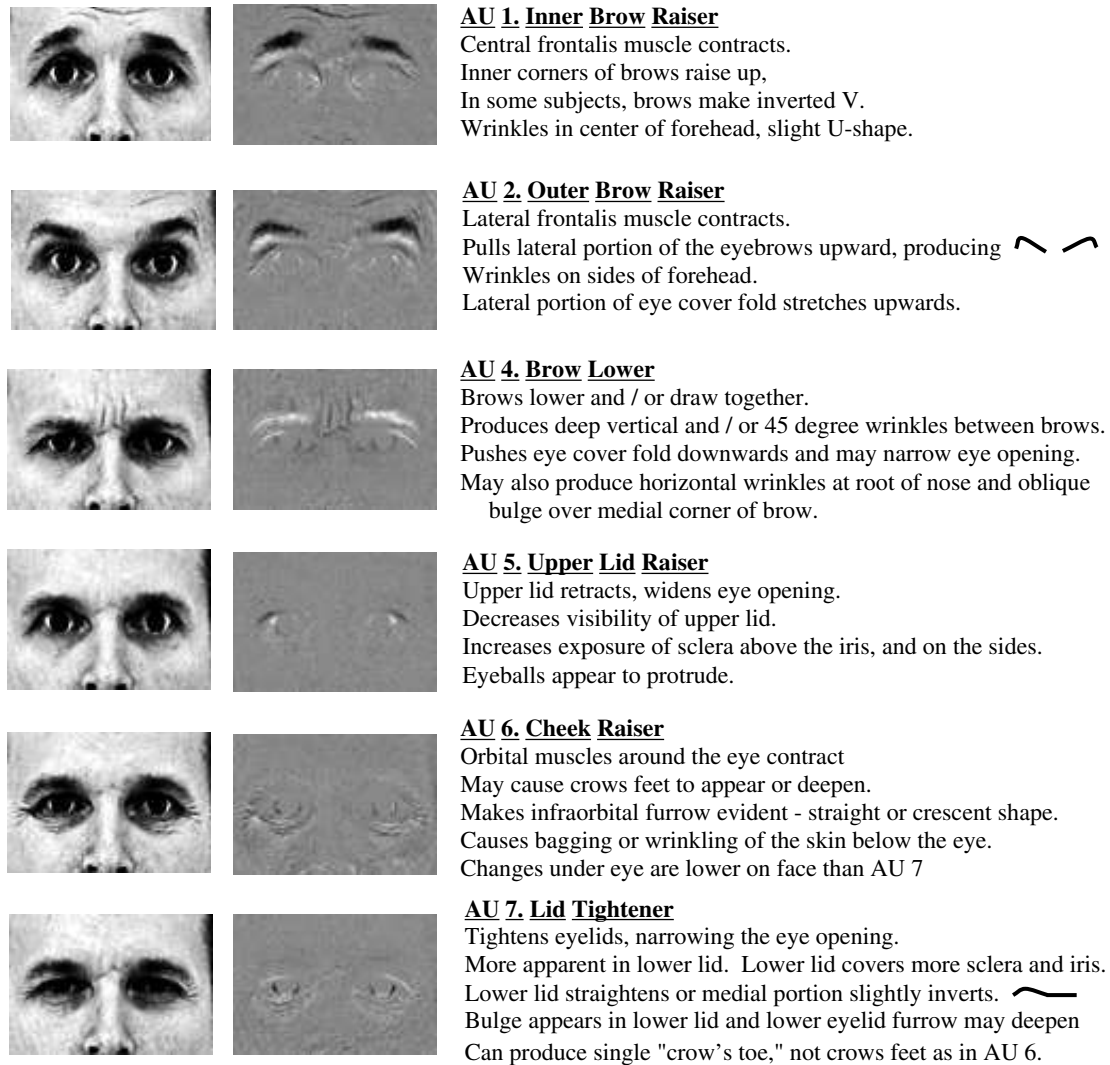


Figure 3.2: Examples of the six actions used in this study. From left to right: cropped image of the action at highest magnitude; difference image obtained by subtracting the neutral image (frame 1 of the sequence); Action unit number; Action unit description adapted from Ekman & Friesen (1978).

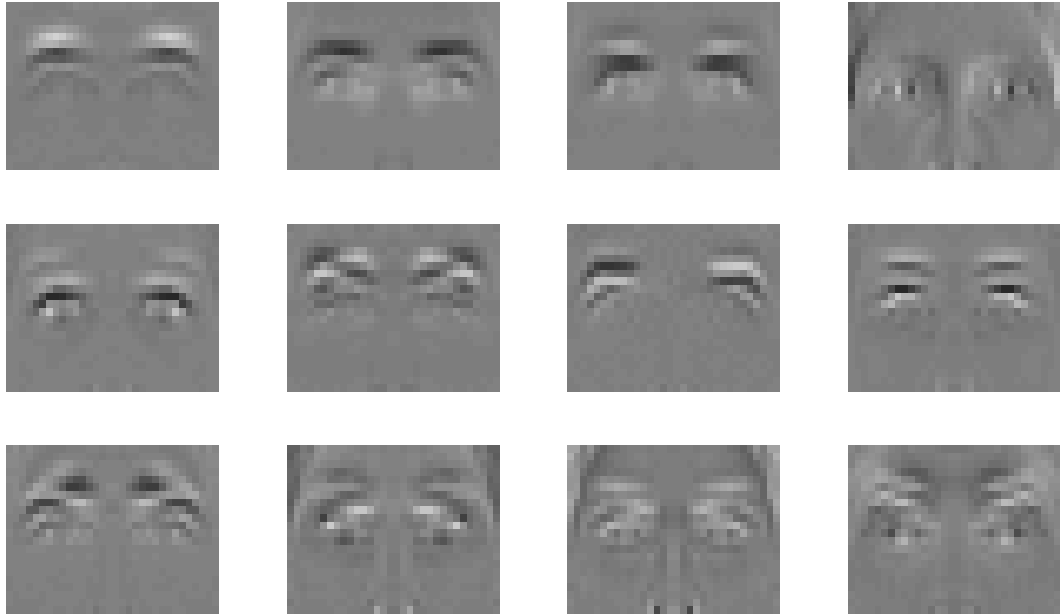


Figure 3.3: First 12 principal components of the dataset of difference images, ordered left to right, top to bottom. The first component appears to code for vertical brow position. The sixth component axis appears to differentiate between AU1, raising the inner corners of the brow, and AU2, raising the 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.

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 [14] 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 3.4a, 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 3.4b 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 the sum squared derivative of the pixel intensities along the segment. This value is estimated by P (Figure 3.4c.) 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 3.4d 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.

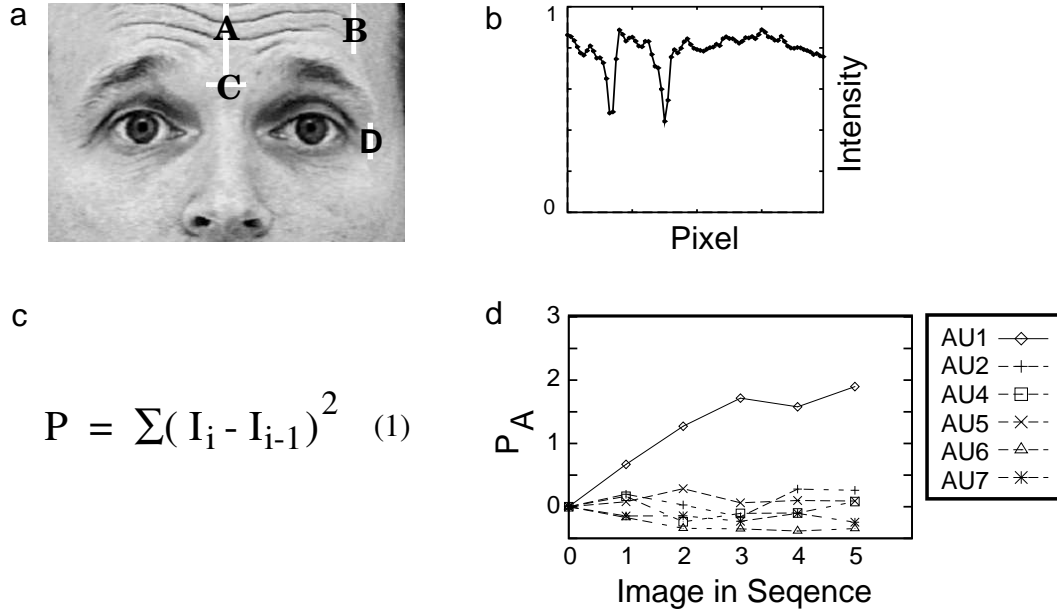


Figure 3.4: a) Wrinkling was measured at four image locations, A-D. b) Smoothed pixel intensities along the line labeled A. c) The wrinkle measure, P . I_i is the intensity of the i th pixel of the segment. Pixel differences approximate the derivative (Jain, Kasturi, & Schunk, 1995). d) P measured at image location A for one subject performing each of the six actions.

Optic flow

Local estimates of motion in the direction of the image gradient were obtained by an algorithm based on the brightness constraint equation [98]:

$$\frac{dI(x, y, t)}{dt} = \frac{\partial x}{\partial t} \frac{\partial I(x, y, t)}{\partial x} + \frac{\partial y}{\partial t} \frac{\partial I(x, y, t)}{\partial y} + \frac{\partial I(x, y, t)}{\partial t} = 0 \quad (3.1)$$

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 smoothed by a 5 x 5 Gaussian kernel. Estimates of



Figure 3.5: 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.

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 3.5. 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 :

$$S(f^n, f^t) = \frac{\sum_i f_i^n \cdot f_i^t}{\sqrt{\sum_i f_i^n \cdot f_i^n} \sqrt{\sum_i f_i^t \cdot f_i^t}} \quad (3.2)$$

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

Naive human subjects

Subjects were nine adult volunteers with no prior knowledge of facial expression measurement. Subjects were provided with a guide sheet similar to Figure 3.2 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 [62]. 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.

Expert coders

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.

3.3.2 Results

Generalization to novel faces was tested using leave-one-out cross-validation [196]. 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 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 3.6 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 3.1 and 3.2.

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 [153] 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, 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.

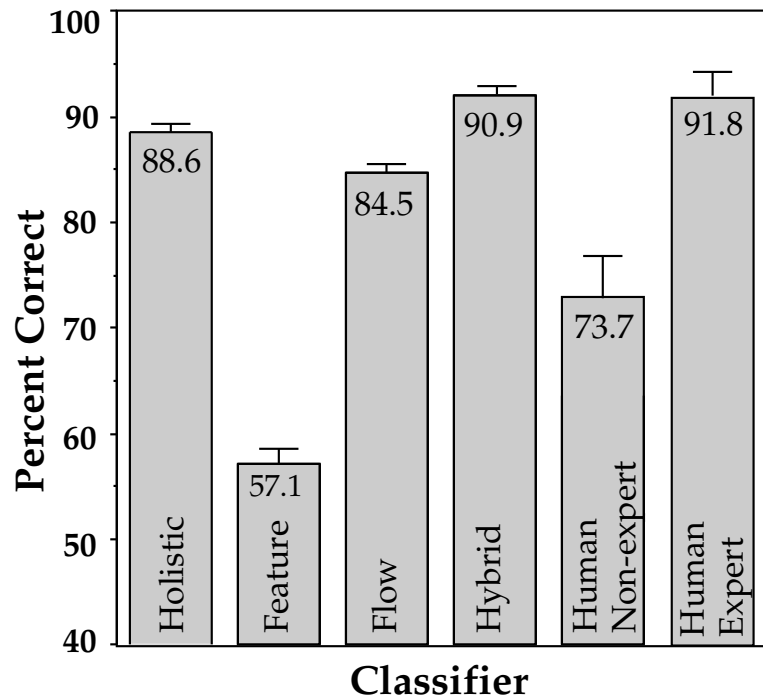


Figure 3.6: Performance comparisons for generalization to novel subjects. Values are percent correct across all test images. Error bars are one standard deviation of the estimate of the success rate in a Bernoulli distribution. Human results were prorated by action and action magnitude to match the proportions in the complete image set.

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 benefitted 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 datapoint 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.

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 ($Z = 7.2, p < 0.001$).

Error analysis

The action confusions made by both naive and expert human subjects are presented in Table 3.1. 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 3.2 shows the action confusions made by the three image analysis systems and the hybrid system. Correlations among the action confusions are given in Table 3.3. 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 [152, 85]. 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.

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

Table 3.1: 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							
	Hol	Mt	Ft	Hyb	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 3.2: Confusion Matrix for the automated classifiers.given action. Hol: Holistic, Mt: Motion, Ft: Feature, Hyb: Hybrid.

	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

Table 3.3: 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***.

3.4 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 92% 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 [153]. 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. [27]), 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 nonfrontal view (eg. [28, 201]).

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 [62]. 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.

Cohn et al. [42] are developing a related system for automatic facial action coding. The Cohn et al. system estimates displacements in a select set of feature points whereas our system captures full field information on skin motion. The Cohn et al. system takes advantage of the precision obtainable through human interaction by manually identifying the more than 40 feature points in the initial image. The system presented here is more automatable since human interaction in our system was limited to the two mouse clicks in the initial image, described above.

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 [15], recognition rates are only just above chance. Likewise, although humans can recognize facial expressions quite well from static graylevel images, expression recognition improves with high temporal resolution video [202]. This system integrates both analysis of surface graylevels and motion information.

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 faces [34, 115]. Our results also suggest that hand-crafted 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 [153] 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-crafted feature measures that performed poorly in this study and others (e.g. [34]). We are presently exploring 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. [58], human interaction and communication (e.g. [66], cognition (e.g. [210], and child development (e.g. [38]. Measurement of observable facial behavior has been combined with simultaneous scalp EEG in the study of physiological patterns associated with emotional states (e.g. [51], and with measures of autonomic nervous system activity to study the relationship of emotion to facial muscles and the autonomic nervous system [65].

Neuropsychological investigations in humans and physiological recordings in primates have indicated a separate neural substrate for recognizing facial expression independent of identity [194, 3, 87], and there is evidence that the recognition of specific facial expressions depends on distinct systems (e.g. [2]). Neural substrates for the perception of two negative emotions, fear and disgust, have recently been differentiated using fMRI [159]. Whereas perception of expressions of fear and anger produced activation in the amygdala [31, 137], perception of disgust in others activated interior insular cortex, an area involved in responses to offensive tastes [207, 109].

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 and two anonymous reviewers for valuable comments on earlier drafts of this paper.

This chapter, in full, is a reprint of material that has been accepted for publication in *Psychophysiology*, Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J., in press, 1998. The dissertation author was primary investigator and primary author of this paper.

Chapter 4

Classifying Facial Actions

4.1 Abstract

The Facial Action Coding System [62], 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 paper 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, which both achieved 96% accuracy for classifying twelve facial actions. The results provide converging evidence for the importance of possessing local filters, high spatial frequencies, and statistical independence for classifying facial actions.

4.2 Introduction

Facial expressions provide information not only about affective state, but also about cognitive activity, temperament and personality, truthfulness, and psychopathology. Facial measurement is used as a behavioral measure of such internal processes in studies of emotion, social interaction, communication, anthropology, personality, and child development (for reviews see [67] [64] [66]). The leading method for measuring facial movement is the Facial Action Coding System [62]. The Facial Action Coding System (FACS) provides an objective means for describing facial signals in terms of component motions, or “facial actions.” FACS is presently performed by highly trained human experts. Recent advances in image analysis open up the possibility of automatic measurement of facial signals. An automated system would make facial ex-

Copyright 1998 IEEE. Reprinted, with permission, from *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Submitted March 25, 1998.

pression measurement more widely accessible as a tool for research and assessment in behavioral science and medicine. Such a system would also have application in human-computer interaction tools. This paper explores and compares methods for classifying Facial Actions in image sequences of faces.

4.2.1 The Facial Action Coding System

The Facial Action Coding System (FACS) was developed by Ekman and Friesen [62] to investigate the relationship between facial behavior and internal state. It was developed in order to study which facial motions are produced under what conditions, how consistently are they produced, and whether the same facial motions are produced under the same conditions across cultures. Examples of the kind of questions that can only be answered by measuring the facial behavior itself include: Are there differences in facial behavior when people are telling the truth as compared to when they are lying? Do different patterns of central nervous system activity accompany different facial movements?

The Facial Action Coding System (FACS) [62] allows precise specification of the morphology (the specific facial actions which occur) and the dynamics (duration, onset and offset time) of facial movement. FACS was developed by determining from palpation, knowledge of anatomy, and videotapes, how the contraction of each of the facial muscles changed the appearance of the face. Ekman and Friesen defined 46 Action Units, or AUs, to correspond to each independent motion of the face. 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 includes the onset, start of apex, end of apex, and offset of each facial action. More than 300 people worldwide have achieved inter-coder agreement on the Facial Action Coding System.

Dynamic information in facial behavior is a critical component of FACS. Timing, for example, contains information for discriminating genuine, emotional smiles from false ones produced voluntarily such as when smiling for the camera [63]. In addition to upward movement of the mouth corners (AU 12), a genuine smile includes an eye scrunch produced by contraction of the outer portion of the sphincter muscle around the eye (AU 6). Genuine smiles typically feature apex coordination, with AU 6 reaching its maximum intensity at the same time as AU12, and are also shorter and smoother in execution than anxious or false smiles. Huang [99] demonstrated that the relative timing of mouth and eye motion in synthesized expressions can shift perception of a smile from a genuine smile, to a voluntary smile, to a sneer.

In recent years a number of studies have appeared showing the rich variety of information that can be obtained by using this tool. Examples include evidence of a facial signal for embarrassment [108]; differences between genuine and simulated pain [49]; cross cultural differences in how infants respond to restraint [39]; and signs of psychopathology [185]. Experiments using FACS, for example, have shown significant differences between the facial signals of suicidal and non-suicidally depressed patients [92]. When responding to the question, “Do you still wish to take your own life?” patients who had recently attempted to kill themselves displayed brief facial signals associated with contempt and disgust, whereas none of the non-suicidal patients displayed those signals. (See [67] for a report of another eighteen studies which measured facial behavior with FACS on these and related topics).

Promising as these findings are, a major impediment remains: 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. Automating the FACS would make it more widely accessible as a research tool, and it would provide a good foundation for human-computer interaction tools. Some success has been achieved for automatic detection of facial actions by tracking the positions of dots attached to the face [93] [105]. A system that detects facial actions from image sequences without requiring application of dots to the subjects face would have much broader utility. Efforts have recently turned to measuring facial actions by image processing of video sequences [14] [42].

Components of FACS have been incorporated into computer graphic systems for synthesizing facial expressions (e.g. *Toy Story* [107], and for parameterizing facial movement [176] [129] [206]. There appears to be some confusion in the computer vision literature between the Facial Action Coding System itself and computer graphic systems that implement aspects of FACS. For example, a number of criticisms of FACS such as ignoring temporal information (cf. [68]), are actually criticisms of some computer graphic implementations of FACS, and not of FACS itself. Another clarification is that 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 motion and image changes they produce in video sequences of face images.

4.2.2 Analysis of facial signals by computer

Recent advances have been made in computer vision for automatic recognition of facial expressions in images. The approaches that have been explored include analysis of facial motion [129] [206] [173] [68], measurements of the shapes of facial features and their spatial arrangements [115], holistic spatial pattern analysis using techniques based on principal component analysis [48] [153] [115] and methods for relating face images to physical models of the facial skin and musculature [129] [193] [121] [68]. A number of methods that have been developed for representing faces for identity recognition may also be powerful for expression analysis. These include Gabor wavelets [50] [113], linear discriminant analysis [23], local feature analysis [156], and independent component analysis [13] [10]. These approaches are reviewed below, but first we present some considerations for building an automatic facial expression analysis system that will be of practical use when presented with the diversity of facial signals that occur during natural communication.

Facial action codes versus emotion categories.

Most systems for automatic facial expression analysis attempt to classify expressions into a few broad categories of emotion, such as happy, sad, or surprised, often quoting behavioral evidence for seven universal facial expressions (see [60] for a review of universals in facial expression). This is a reasonable starting point, but there tends to be a basic misunderstanding of these studies in the computer vision literature. The universal expression studies demonstrate consistent facial signals across cultures for seven discrete emotions, but do not imply that all facial expressions are subsumed by one of these categories.

Real facial signals consist of thousands of distinct expressions, for many of which a gross category assignment would be impossible, misleading, or simply insufficient [83]. Signals of two or more emotions may occur in the same facial expression, such as blends of happiness and disgust which produce "smug," or blends of happiness and sadness which produce "nostalgia" [61]. For applications such as user-interfaces, computer games, or TV ratings, it may be important to distinguish "happy-surprise" from "fearful-surprise" or "horrified-surprise" which implies disgust as well. Within an emotion category such as anger, there are variations in the intensity of the emotion, such as annoyance and fury, and variants of the emotion such as vengeance and resentment. Most facial movements shown during a social interaction are not relevant to emotion at all, but are what Ekman called "conversational signals" [55]. A variety of facial movements provide different kinds of emphasis to speech and can also provide information about syntax. This information is important for understanding the nature of conversation, identifying fluctuations in the speaker's level of involvement with what he or she is saying, and could augment a speech recognition system. Finally, there are facial tics, mannerisms, and other peculiarities in facial movement which may be relevant to personality, psychopathology, and/or brain lesions.

If automated facial measurement were, then, to be constructed simply in terms of seven elementary emotional categories, much important information would be lost: blends of two emotions, variations within an emotional category, variations in intensity, conversational signals, and idiosyncratic facial movements. Such information need not be lost if the automated system is based on the Facial Action Coding System, which provides a description of the basic elements of any facial movement, analogous to phonemes in speech. Facial Action codes provide a rich description of facial behavior, and a system that outputs Facial Action codes instead of, or in addition to, emotion category labels will be a more powerful tool for applications both in industry and behavioral science.

For basic science inquiries into the relationship of facial behavior and internal state, objective measurements of facial behavior are required and a system that only produced emotion category labels could not be used for that purpose. Some computer vision systems do provide explicit descriptions of facial movement [129] [206][68], but it is not known whether these descriptions are sufficient for describing the full range of facial behavior. These descriptions are not readily interpretable in terms of Facial Action codes, and there is not yet any empirical data establishing relationships between these motion descriptions and internal state. A large body of behavioral data already exists establishing the relationship of Facial Action codes to emotions, emotion intensity, variations, blends, and conversational signals.

Feigned expressions of emotion.

Another important issue for developing an automated facial expression analysis system is obtaining a reliable set of training images from which to build the system [83]. Most systems have utilized datasets of voluntary expressions, in which subjects were asked to "look happy," "sad," or "surprised." While the use of such feigned expressions is again a reasonable starting point, there are limitations to such a datasets that must be recognized. Voluntary expressions differ from spontaneous expressions in response to actual emotion. There is evidence that they are mediated by different neural substrates [168]. When subjects produce voluntary expres-

sions, some actions tend to be exaggerated, some actions can be incorrectly included, while some actions that are usually present when the emotion is experienced are omitted [55]. There is a tendency to exaggerate facial motions that are used as conversational signals, such as a full brow raise, and omit other motions for which we have less practice at cognitive control during speech, such as many signals in the eye region [61]. It can be important to detect these latter signals, since they are also less successfully masked or neutralized when attempting to hide one's emotional state [61].

More reliable datasets could be obtained a number of ways, including recording images of subjects as they watch emotive videos. FACS provides another means of obtaining a reliable dataset. Subjects can be instructed to perform specific Facial Actions, and the veracity of the dataset would not depend on the subjects' acting abilities, but rather on the FACS score describing the actual facial behavior in the images. Figure 4.1 shows image sequences from our database of two subjects performing an individual Facial Action, the inner brow raiser.



Figure 4.1: Example action sequences from the database. The example sequences show two subjects demonstrating AU1 starting from a null expression and ending with a high magnitude example of the action. Frame 2 is a low magnitude example of the action, frames 3 and 4 are medium magnitude examples, and frames 5 and 6 are high magnitude.

Analysis of facial motion.

The majority of work on facial expression recognition has focused on facial motion analysis through optic flow estimation. In an early exploration of facial expression recognition, Mase [129] used optic flow to estimate the activity in 12 of the 44 facial muscles. For each muscle he defined a window in the face image and an axis along which each muscle expands and contracts. The mean similarity of the flow vectors inside the window to this axis provided a coarse estimate of the activity of the muscle. Mase also explored a statistically driven technique for recognizing facial expressions from optic flow. Means and covariances of optic flow in local regions of the face provided a high-dimensional feature vector, of which the 15 measures with the highest ratios of between-class to within-class variability were used for expression classification.

Yacoob & Davis [206] combined tracking of major facial features with local analysis of optic flow. They constructed a mid-level representation of facial motion by first locating and tracking prominent facial features, and then quantizing the optic flow within subregions of each feature into eight principal directions. The mid-level representation consisted of such descriptions as “right mouth corner raises.” These descriptions were then classified into one of six facial expressions using a set of heuristic rules.

Rosenblum, Yacoob & Davis [173] 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 similar to [206], 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. The output of each expression network over the full time-course of the expression was then analyzed, and a set of heuristic rules were established to determine whether to accept or reject the image sequence as an example of that expression.

Cohn et al. [42] are developing a system for facial action coding based on human-computer interaction. Over 40 feature points were manually located in the initial face image, and the displacements of these feature points were estimated by optic flow. Discriminant functions were employed to classify facial actions from the set of 40 displacements.

Model-based techniques.

Several facial expression recognition systems have employed explicit physical models of the face [129] [193] [121] [68]. Mase’s approach described above, for example, employed an extremely simple physical model. Essa & Pentland [68] extended a detailed anatomical and physical model of the face developed by Terzopoulos and Waters [193] applied it to both recognizing and synthesizing facial expressions. The model included 44 facial muscles, their points of attachment to the skin, and the elastic properties of the skin modeled in a geometric mesh. 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 in a recursive estimation and control framework, and the estimated forces were used to classify the facial expressions. In a variation on this work, Essa and Pentland generated templates of 2-D motion energy by back-projecting the “corrected” motion into the 2-D image. Facial expressions were recognized by template matching in the 2-D image space, bypassing the more time consuming physical model.

In a model-based system, classification accuracy is limited by both the accuracy of mapping image onto the model and the validity of the model itself. There are numerous factors that influence the motion of the skin following muscle contraction, and it would be difficult to accurately account for all of them in a deterministic model. In this paper 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.

Beymer, Shashua, and Poggio [28] demonstrated the potential of example-based ap-

proaches for analysis and synthesis of face images. They 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.

Feature-based approaches.

One of the earliest techniques for recognizing facial identity in images was based on the computation of a set of geometrical features of the face such as nose length, chin shape, and distance between the eyes [106] [34]. Geometrical features relevant to facial expression analysis might include mouth shape, eye to eyebrow distance, or the magnitude of wrinkles in specified locations on the face. A difficulty with feature-based approaches to image analysis is that the specific image features relevant to the classification may not be known in advance, and the selected set of features may fail to capture sufficient discriminative information.

An alternative to feature-based image analysis is template matching. Templates capture information about configuration and shape that can be difficult to parameterize. Template matching has been shown to outperform feature-based methods for face recognition [34] [115]. In previous work on this project [14], we explored classification of upper facial actions using feature measurements such as the increase of facial wrinkles in specific facial regions and a measure of eye opening. We found that these feature measurements were less reliable indicators of facial actions than template-based classifiers. The poor performance of the feature measures was attributed to differences in patterns of facial wrinkling across subjects due not only to age but to variations in facial morphology as well.

Feature-based and template-based methods need not be mutually exclusive. Lanitis, Taylor, & Cootes, [115] 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 this set of features with parameters containing information about modes of variation in graylevel images, using the principal component analysis techniques described in Section 4.2.2.

Holistic analysis.

One form of holistic analysis is to train a neural network through back-propagation to classify facial expressions directly from the image pixels. Kobayashi, Tang, and Hara [111] obtained 90% accuracy using a neural network to recognize six basic expressions from selected columns of the graylevel image pixels. Neural network approaches to image analysis can be advantageous for face processing since the physical properties relevant to the classification need not be specified in advance. The weights are equivalent to a learned set of templates. The network learns the relevant “features” from the statistics of the image set, where these features could be either local or involve relationships among multiple image locations. This is particularly useful when the specific features relevant to the classification are unknown. (See [200] for a review of connectionist approaches to processing images of faces.).

Another holistic spatial representation is based on the principal components of the image pixels [47] [197]. The principal components are the eigenvectors of the pixelwise covariance matrix of the set of images. A low-dimensional representation of the face images with minimum reconstruction error is obtained by projecting each image onto the first few principal component axes. Accurate alignment of the faces is critical to the success of such image-based approaches. Principal components analysis has been applied successfully to recognizing both facial identity [47] [197], and facial expressions [48] [153] [14]. Feedforward networks taking such holistic representations as input can also successfully classify gender from face images [48] [78]. A class-specific linear projection of a principal components representation has recently been shown to give highly accurate identity recognition performance when other examples of the same faces are available to calculate the projection weights [23]. New views of a face can be synthesized from a sample view using principal components representations of face shape and texture [201]. Principal component representations of face images have also been shown to account well for human perception of distinctiveness and recognizability [152] [85].

Representations based on principal components analysis address the second-order statistics of the image set, and do not address high-order statistical dependencies such as the relationships among three or more pixels. In a task such as facial expression recognition, much of the important information may be contained in the high-order relationships among the image pixels. Independent component analysis (ICA) is a generalization of PCA which uses the high-order moments of the input in addition to the second-order moments. An unsupervised learning algorithm for performing ICA was recently developed [24]. This algorithm has proven successful for separating a set of randomly mixed auditory signals, known as the cocktail party problem [24], and also has been applied to separate EEG signals [128], fMRI signals [131], and find image filters that give independent outputs from natural scenes [25]. A representation for face recognition based on the independent components of face images has recently been developed [13] [10], and was found to be superior to the PCA representation for classifying facial identity across changes in expression and changes in pose.

Penev and Atick [156] recently developed a topographic representation based on principal component analysis. The representation is based on a set of kernels that are optimally matched to the second-order statistics of the input ensemble. The kernels were obtained by performing zero phase whitening of the principal components, followed by a rotation to topographic correspondence with pixel location. Penev and Atick call the technique local feature analysis (LFA) because the resulting kernels contain spatially local regions of nonzero value, but in this paper we class this technique as holistic, as with ICA, since the image-dimensional kernels result from statistical analysis over the whole image. Atick's group obtained the highest recognition performance so far on the FERET face recognition test [160]. Although the actual techniques employed during this test have not been disclosed, it has been implied that they included some of the principles embodied in LFA.

Local spatial analysis

Representations based on local spatial filters may be superior to spatially global representations for image classification. Padgett & Cottrell [153] improved the performance of their

facial expression recognition system by performing PCA on 32×32 subimage patches of the face images, and using these 32×32 principal component images as convolution kernels. This finding is supported by Gray, Movellan & Sejnowski [79] who also obtained better performance for visual speechreading using the principal components of subimage patches as convolution kernels over a representation based on the principal components of the full images.

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 [50]. Representations based on the outputs of these filters at multiple spatial scales, orientations, and spatial locations, have been shown to be useful for recognizing facial identity [113]. In a direct comparison of face recognition algorithms, Gabor filter representations gave better face recognition performance than representations based on principal components analysis [212].

4.3 Overview of Approach

This paper explores and compares methods for classifying facial actions in image sequences of faces. We examine a number of techniques that have been presented in the literature for processing images of faces, and compare their performance on this image analysis task.

1. Optic Flow. Two methods for estimating optic flow were implemented and compared: a fast gradient-based method for calculating flow between pairs of images based on [98], and a correlation-based method [184]. Local smoothing is commonly imposed on flow fields to clean up the signal. We also examined the effects of local smoothing on classification of facial motion.

2. Holistic analysis. Next, we examined classification performance based on holistic spatial analysis of the graylevel images. We compared four holistic representations: principal component analysis (PCA), which finds a set of image-dimensional kernels that decorrelate the second order statistics of the dataset; independent component analysis (ICA), which minimizes the higher order dependencies in addition to the covariance; Local feature analysis (LFA), which is a topographic representation based on principal component analysis; and Fisher’s linear discriminants (FLD), which computes a class-specific linear projection of the PCA representation onto lower dimensions.

3. Local filters. In addition, we examined representations based on the outputs of local filters. A local filter consisting of a simple 15×15 Gaussian provided a benchmark for classification performance using local filters. Three local filters based on principal components analysis were then compared. The first filter consisted of the principal component eigenvectors of 15×15 subimage patches taken from *random* locations throughout the images. These principal component eigenvectors were used as convolution kernels for filtering the entire face image. The second local representation consisted of the principal components of 15×15 subimage patches at *fixed* locations in each image. These PCA-based local representations were compared to a Gabor wavelet representation for classifying facial action. The local-PCA *random* representation is related to the Gabor representation in that it provides the local amplitude spectrum of the image. In order to further compare the Gabor and local-PCA representation, we devised an intermediate representation based on local PCA that contained multiscale, hierarchical properties corresponding to the Gabor filter bank. This representation matched the Gabor representation for both spatial scale

and number of filters.

4. Human Subjects. The ability of naive human subjects to classify facial actions in the same images that were presented to the classification algorithms provided a benchmark for the performance of those systems. Since the long-term goal of this project is to replace human expert coders with an automated system, a second benchmark was provided by the agreement rate of expert coders on these images.

4.4 Image Database

We collected a database of image sequences of subjects performing specified facial actions. The full database contains over 1100 sequences containing over 150 distinct actions, or action combinations, and 24 different subjects. Trained FACS experts provided demonstrations and instructions to subjects on how to perform each action. The selection of images was based on FACS coding of stop motion video. The images were coded by three experienced FACS coders certified with high intercoder reliability. The criterion for acceptance of images was that the requested action and only the requested action was present. 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 4.2 for a summary of the actions examined. 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 [62] which allowed us to treat them separately.

We obtained examples of the 12 actions from 20 subjects. Each example action in the database consisted of a sequence of six images, beginning with a neutral expression and ending with a high magnitude¹ muscle contraction (see Figure 4.1).

The first image in each sequence was located by manually marking three points: the centers of the eyes and mouth. The performance of some methods depended on obtaining highly accurate normalization of the faces. The three coordinates were then used to center, rotate, scale, and finally crop a window of 60×90 pixels around the region of interest (eyes or mouth). To control the variation in lighting between frames of the same sequence and in different sequences, we applied a logistic filter whose parameters were chosen to match the statistics of the grayscale levels of each sequence [138]. This procedure enhanced the contrast, performing a partial histogram equalization on the images.

4.5 Optic Flow Analysis

We compared two methods for calculating flow, a fast gradient-based method for calculating flow between pairs of images based on [98], and a correlation-based method [184]. We also examined the contribution of local smoothing of the flow signal to action classification.

¹The term “magnitude” replaces the term “intensity” used in FACS to avoid confusion with image intensity.


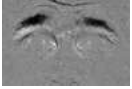



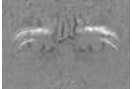

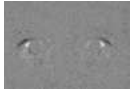
















Upper Face		Action Unit	Subjects
		1 Inner brow raiser	9
		2 Outer brow raiser	10
		4 Brow lowerer	18
		5 Upper lid raiser	20
		6 Cheek raiser	5
		7 Lid tightener	18
Lower Face			
		17 Chin raiser	8
		18 Lip puckerer	4
		9 Nose wrinkler	4
		25 Lips part	
		10 Upper lip raiser	5
		25 Lips part	
		16 Lower lip depressor	4
		25 Lips part	
		20 Lip stretcher	6
		25 Lips part	

Figure 4.2: 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, Action Unit name, and number of subjects imaged.

4.5.1 Gradient-based optic flow

Local estimates of motion in the direction of the image gradient were obtained from pairs of images by an algorithm based on the intensity conservation equation [98]:

$$\frac{dI(x, y, t)}{dt} = 0 \Rightarrow \frac{\partial x}{\partial t} \frac{\partial I(x, y, t)}{\partial x} + \frac{\partial y}{\partial t} \frac{\partial I(x, y, t)}{\partial y} + \frac{\partial I(x, y, t)}{\partial t} = 0 \quad (4.1)$$

where $\vec{u} \triangleq \frac{\partial x}{\partial t}$ and $\vec{v} \triangleq \frac{\partial y}{\partial t}$ are the local velocities in the x and y directions. This equation assumes that there is no overall gain or loss of brightness in the image over time, and that any local changes in brightness can be accounted for by shifts in spatial position. The image velocity is defined in terms of the spatial and temporal gradients of the image, $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, and $\frac{\partial I}{\partial t}$. These image gradients can be estimated directly from the pair of images, and we can solve for \vec{u} and \vec{v} to obtain an estimate of the velocity in the direction of the image gradient.

Images were smoothed by a 5x5 Gaussian kernel. Estimates of the spatial gradients, Δx and Δy , were obtained with horizontal and vertical Sobel edge filters. The temporal gradient was estimated by $\Delta t = I(x, y, t_1) - I(x, y, t_0)$. We took as our local estimates of image velocity

$$\hat{u} = \frac{\Delta t}{\Delta x}, \quad \hat{v} = \frac{\Delta t}{\Delta y} \quad (4.2)$$

Gradient-based techniques for estimating optic flow such as this one give reliable estimates of velocity only at points where the spatial and temporal gradient of the image sequence is high. We therefore retained only the velocities from those locations. Velocity estimates were set to zero at locations at which the total edge measure $r = \Delta x^2 + \Delta y^2$ was beneath a threshold of 0.2. One of the advantages of this simple local estimate of flow was speed. It took 0.25 seconds on a Dec Alpha 2100a processor to compute one flow field.

4.5.2 Correlation-based optic flow

The second estimate of optic flow was obtained by correlation-based extraction of local velocity information [184]. As in the gradient-based approach, this algorithm assumes a luminance conservation constraint. 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 are computed, thus giving the matching strength, or *response*, at each pixel in the search window \mathcal{W}_s .

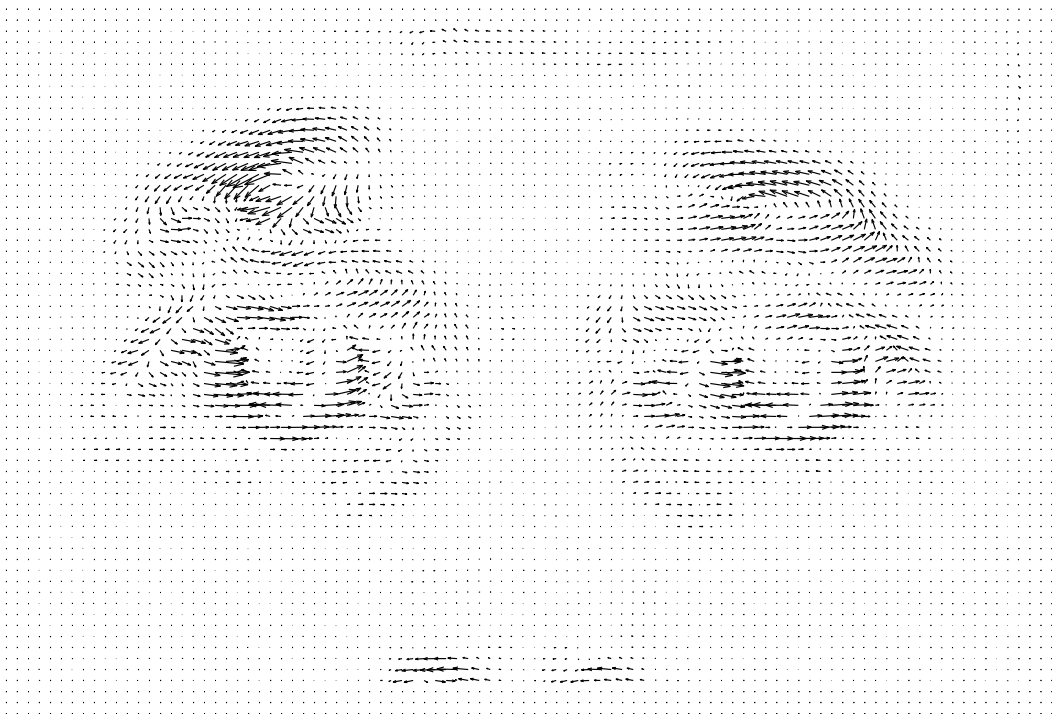


Figure 4.3: Optic flow for AU1 extracted using local velocity information extracted by the correlation-based technique, with no spatial smoothing.

At the end of this process \mathcal{W}_s is covered by a response distribution \mathcal{R} in which the response at each point gives the frequency of occurrence, or likelihood, of the corresponding value of velocity. Velocity is then estimated using the weighted least squares estimate in (4.3). Figure 4.3 shows an example flow field obtained by this algorithm.

$$\hat{u} = \frac{\sum_u \sum_v \mathcal{R}(u, v) u}{\sum_u \sum_v \mathcal{R}(u, v)} \quad \hat{v} = \frac{\sum_u \sum_v \mathcal{R}(u, v) v}{\sum_u \sum_v \mathcal{R}(u, v)} \quad u, v \in [-2, 2] \quad (4.3)$$

4.5.3 Local smoothing

To refine the conservation constraint estimate $\mathcal{U}_{cc} = (\hat{u}, \hat{v})$ obtained above, we estimate the velocity at each pixel \mathcal{P} from the velocities in a neighborhood around \mathcal{P} . Assuming no spatial discontinuities in the motion, the velocity $\mathcal{U}'_{cc} = (\hat{u}', \hat{v}')$ at each \mathcal{P}' in a neighborhood of \mathcal{P} can be thought of as a measurement of the velocity of \mathcal{P} . The local neighborhood estimate of velocity, $\overline{\mathcal{U}}$, is a weighted sum of the velocities at \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, 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:

$$\mathcal{U} = \arg \min \{ \|\mathcal{U} - \mathcal{U}_{cc}\|_{\mathcal{S}_{cc}} \bigwedge \|\mathcal{U} - \overline{\mathcal{U}}\|_{\overline{\mathcal{S}}} \} \quad (4.4)$$

Since we do not know the *true* velocity, this estimate must be computed iteratively. To update the field we use the equations [184]:

$$\begin{aligned} \mathcal{U}^0 &= \mathcal{U}_{cc} \\ \mathcal{U}^{k+1} &= [\mathcal{S}_{cc}^{-1} + \overline{\mathcal{S}}^{-1}]^{-1} [\mathcal{S}_{cc}^{-1} \mathcal{U}_{cc} + \overline{\mathcal{S}}^{-1} \overline{\mathcal{U}}^k] \end{aligned} \quad (4.5)$$

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

4.6 Holistic Analysis

The holistic spatial analysis algorithms each found a set of n -dimensional data-driven image kernels, where n is the number of pixels in each image. The analyses were performed on the difference (or δ) images (Figure 4.2), 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 [138]. Holistic kernels for the upper and lower-face subimages were calculated separately.

We began with the zero-mean data matrix X where the δ -images were stored as row vectors x_j :

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{12} & \dots & x_{2n} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad (4.6)$$

In the following descriptions, n is the number of total 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.

4.6.1 Principal Component Analysis: “EigenActions”

This approach is based on [48] and [197], with the primary distinction in that we performed principal components 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 :

$$S = PDP^T \quad (4.7)$$

Examples of the eigenvectors are shown in Figure 4.4. 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.



Figure 4.4: First 8 principal components of the difference images for the upper face actions (top), and lower face actions (bottom). Components are ordered left to right, top to bottom.

4.6.2 “FisherActions”

This approach is based on the original work by Belhumeur and others [23] that showed that a class-specific linear projection of a principal components representation of faces improved identity recognition performance. The method is based on Fisher’s linear discriminant (FLD) [71], which projects the images into a subspace in which the classes are maximally separated. 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 [180] [84]. 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 δ -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 $\chi \triangleq \{\chi_1, \chi_2, \dots, \chi_c\}$ be the set of all $N = |\chi|$ data, divided into c classes. Each class χ_i is composed of a variable number of images $x_i \in \mathbb{R}^n$. The between-class scatter matrix S_B and the inter-class scatter S_W are defined as

$$S_B \triangleq \sum_{i=1}^c |\chi_i| (\mu_i - \mu)(\mu_i - \mu)^T \quad \text{and} \quad S_W \triangleq \sum_{i=1}^c \sum_{x_k \in \chi_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (4.8)$$

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

$$W_{opt} = \arg \max_W J(W) \triangleq \arg \max_W \frac{\det(W^T S_B W)}{\det(W^T S_W W)} = \{w_1, w_2, \dots, w_{c-1}\} \quad (4.9)$$

The $\{w_i\}$ are the solution of the generalized eigenvalues problem

$$S_B w_i = \lambda_i S_W w_i \quad \text{for} \quad i = 1, \dots, c-1 \quad (4.10)$$

Performing PCA on the total scatter matrix $S_T = S_W + S_B$ to project the feature space to \mathbb{R}^{N-c} greatly simplifies the calculations [23]. Calling W_{pca} the matrix of the transformation,

$$W_{pca} \triangleq \arg \max_W |W^T S_T W| \quad (4.11)$$

we can project S_W and S_B :

$$\tilde{S}_B \triangleq W_{pca}^T S_B W_{pca} \quad \text{and} \quad \tilde{S}_W \triangleq W_{pca}^T S_W W_{pca} \quad (4.12)$$

The original FLD problem can thus be reformulated as:

$$W_{fld} = \arg \max_W J(W) \triangleq \arg \max_W \frac{\det(W^T \tilde{S}_B W)}{\det(W^T \tilde{S}_W W)} = \{w'_1, w'_2, \dots, w'_{c-1}\} \quad (4.13)$$

From 4.9 and 4.13, $W_{opt} = W_{pca} W_{fld}$, 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 now full-rank.

4.6.3 Independent Component Analysis

Representations such as "Eigenfaces" [197], "Holons" [47] and "Local Feature Analysis" [156] are based on principal components analysis and are optimally matched to the second-order statistics of the image set, the pixelwise covariances, but are insensitive to the high-order statistics of the image set. 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 provides an image representation that is sensitive to all of the statistical dependencies in the image set, not just the covariances.

The independent component representation was obtained by performing "blind separation" on the set of face images [13] [10]. The independent components of the face images were separated according to the image synthesis model of Figure 4.5. The δ – 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 , which produced statistically independent outputs, U .

The ICA filters, W , were found using an unsupervised learning algorithm derived from the principle of optimal information transfer through sigmoidal neurons [24]. The algorithm maximizes the mutual information between the input and the output of a logistic transfer function, which is equivalent to maximizing the entropy of the output. 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 is the same as the cumulative density functions of the underlying independent components (up to a scaling and translation) it can be shown that maximizing the mutual information between the input and output also minimizes the mutual information between the individual outputs [141] [25].

The ICA outputs, U , provided a set of independent basis images for the expression images. These basis images can be considered a set of statistically independent image features, where the pixel values in each feature image were statistically independent from the pixel values in the other feature images. The ICA representation consisted of the coefficients, \mathbf{b} , for the linear combination of independent basis images, \mathbf{u} , that comprised each face image \mathbf{x} (Figure 4.5, Bottom). These coefficients were obtained from W^{-1} (see [10]).

Examples of the independent components of the expression images are shown in Figure 4.6. The ICA basis images were spatially 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 & Sejnowski, 1997).

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 \bar{x} be the overall mean of a coefficient, and \bar{x}_j be the mean for person j . The ratio of between-class to within-class variability, r , for each coefficient is defined as

$$r = \frac{\sigma_{between}}{\sigma_{within}} \quad (4.14)$$

Where $\sigma_{between} = \sum_j (\bar{x}_j - \bar{x})^2$ is the variance of the j class means, and $\sigma_{within} = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ is the sum of the variances within each class. The first p components selected

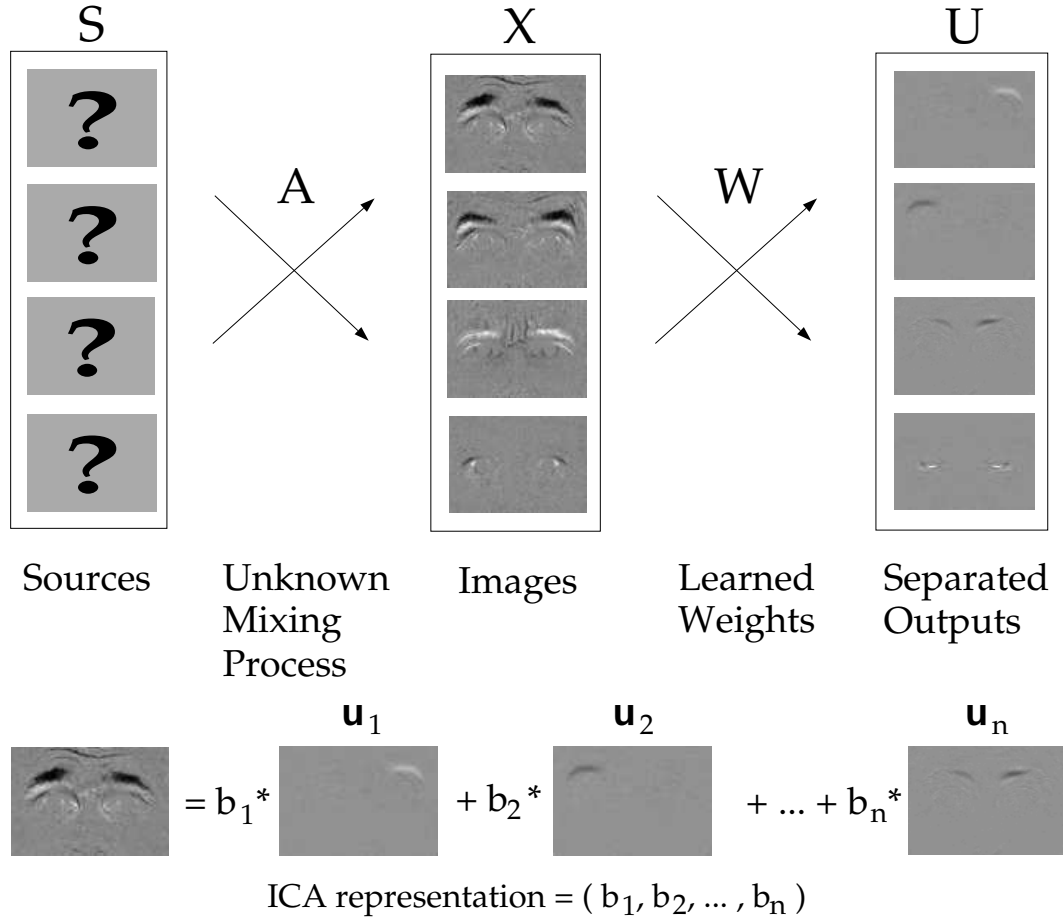


Figure 4.5: TOP: Image synthesis model for the ICA representation. BOTTOM: The ICA representation.

by class discriminability comprised the independent component representation.

4.6.4 Local Feature Analysis (LFA)

Local Feature Analysis (LFA) defines a set of topographic, local kernels that are optimally matched to the second-order statistics of the input ensemble [156]. The kernels are derived from the principal component axes, and consist of "sphering" the PCA coefficients to equalize their variance [7], followed by a rotation to pixel space.

As in the global PCA representation (Section 4.6.1), we begin with the zero-mean matrix of δ -images, X (4.6), and calculate the principal component eigenvectors P according to $S = PDP^T$. Penev & Atick [156] defined the following set of topographic kernels based on P and D , where "topographic" indicates that the kernels are indexed by spatial location:

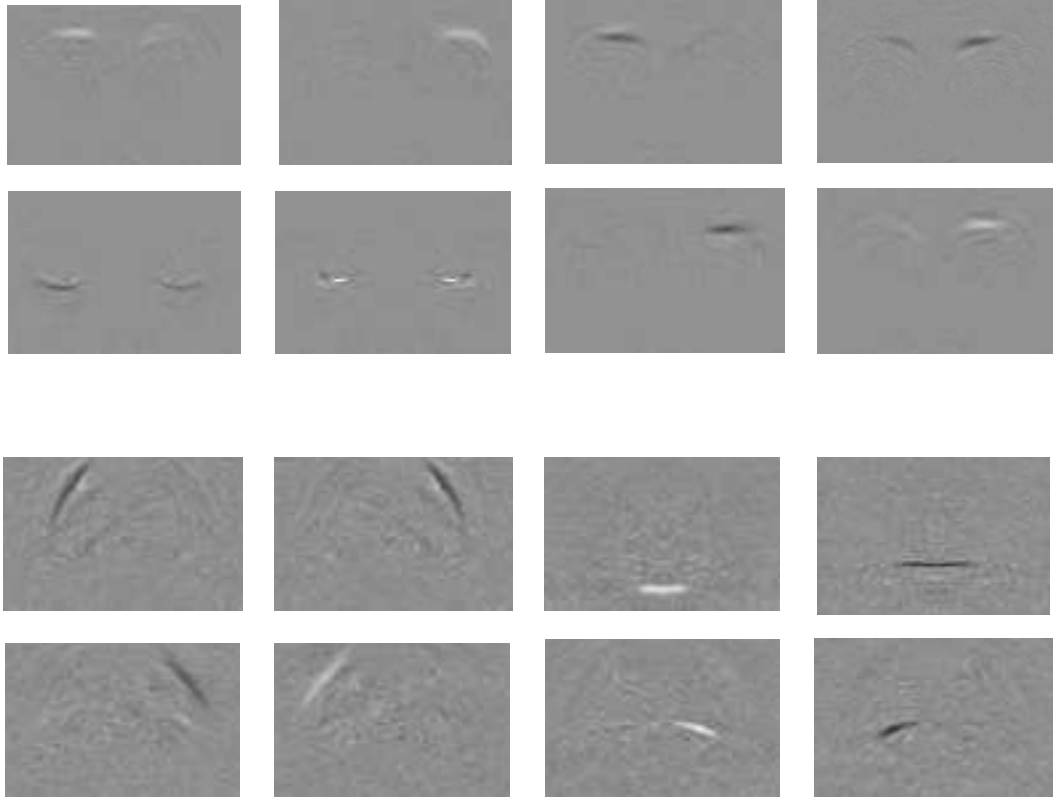


Figure 4.6: Example independent component images of the upper face (top) and lower face actions (bottom).

$$K = PVP^T \quad \text{where} \quad V = D^{-\frac{1}{2}} = \text{diag}\left(\frac{1}{\sqrt{\lambda_i}}\right) \quad i = 1, \dots, p \quad (4.15)$$

where λ_i are the eigenvalues of S . The rows of K contain kernels with spatially local properties. Note that the matrix V is the inverse square root of the principal components covariance matrix. This transform spheres the principal component coefficients (normalizes their output variance to unity) and minimizes correlations in the LFA output. The kernel matrix K transforms X to the LFA output O (see Figure 4.7.)

$$O = KX^T = \begin{bmatrix} O(x_1) \\ O(x_2) \\ \vdots \\ O(x_N) \end{bmatrix} \quad O(x_j) \in R^{1 \times n} \quad (4.16)$$

The original images X can be reconstructed from O by $X^T = K^{-1}O$.

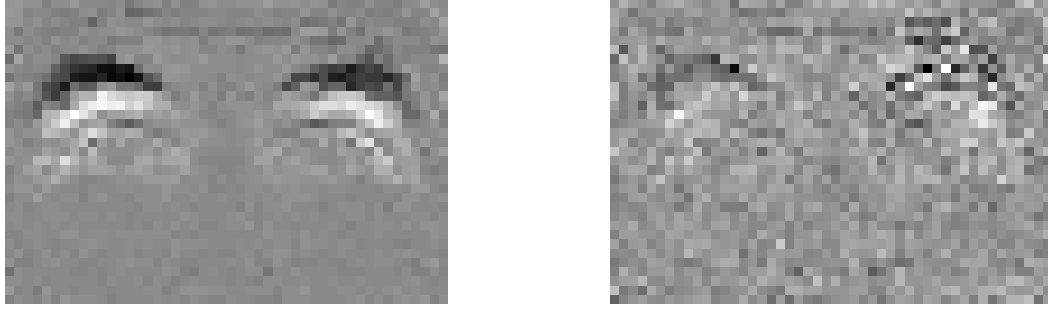


Figure 4.7: An original δ -image on the left, with its corresponding LFA representation $O(x)$ on the right.

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 $p \ll n$ linearly independent variables, there are residual correlations in the output. Penev & Atick proposed 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} .

Penev & Atick [156] presented methods for image *representation*, but did not address the application of local feature analysis to image *recognition*. The sparsification algorithm proposed by Penev & Atick selected a different set of points, \mathcal{M} , for each image, which is problematic for recognition. In order to make the representation amenable to recognition, we diverged from their sparsification algorithm by finding a single set of \mathcal{M} points for all images. At each time step, the point with the largest mean reconstruction error *across all of the images* was added to \mathcal{M} .

At each step, the point added to \mathcal{M} is chosen as

$$\arg \max \langle \|O - O^{rec}\|^2 \rangle \quad (4.17)$$

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

$$\mathcal{Y} = \beta \mathcal{X} \quad (4.18)$$

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:

$$\beta = \frac{\mathcal{Y} \mathcal{X}}{(\mathcal{X}^T \mathcal{X})} = \frac{(O^{rec})^T O(\mathcal{M}, N)}{O(\mathcal{M}, N)^T O(\mathcal{M}, N)} \quad (4.19)$$

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

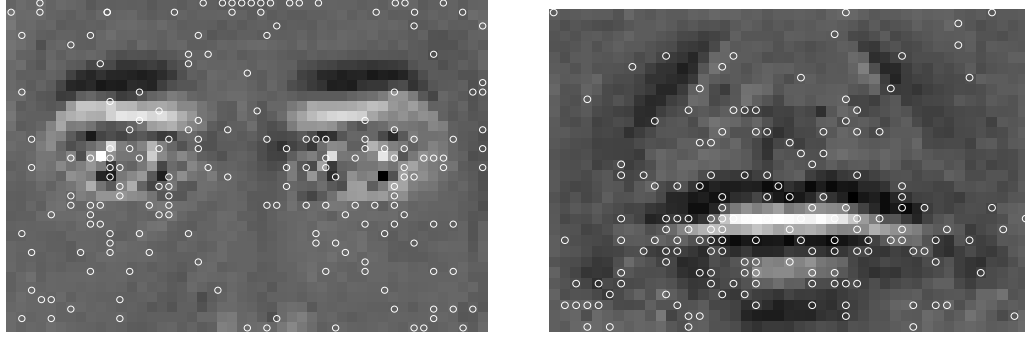


Figure 4.8: The first 155 points selected by the sparsification algorithm superimposed to the mean images of the upper and lower face actions.

Equation 4.19 can also be expressed in terms of the correlation matrix of the outputs, $C = O^T O$, as in [156]:

$$\beta = C(\mathcal{M}, N)C(\mathcal{M}, \mathcal{M})^{-1} \quad (4.20)$$

The termination condition can be either $|\mathcal{M}| = N$ or $\langle \|O - O^{rec}\|^2 \rangle \leq \varepsilon$. We calculated all possible N points and then tested for optimal classification performance. Figure 4.8 shows the locations of the points selected by the sparsification algorithm. The algorithm converges to the least-squares minimum error. If K is calculated retaining all eigenvectors then the reconstruction is perfect and the error for $|\mathcal{M}| = N$ is exactly zero.

4.7 Local Representations

The approaches described so far were all “global” meaning that the kernels for the representation were derived from the entire image. We explored five different kinds of local representations based on filters that act on small spatial regions within the images. Three of the filters are based on PCA, whereas one relies on a biologically inspired wavelet decomposition.

4.7.1 Gaussian kernel

A simple benchmark for the local filters consisted of a single Gaussian kernel. δ – Images were convolved with a 15×15 Gaussian kernel and the output was downsampled by 0.25. The dimensionality of the final representation was $\frac{n}{4}$.

4.7.2 Local PCA: Random patches

A local representation based on the principal components of subimage patches (local PCA) outperformed the representation based on the principal components of the full images (global

PCA) for classifying facial expressions [153]. Local basis functions were obtained from the principal component eigenvectors of image patches selected from random image locations. These results were supported by [79] for lipreading. 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 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 4.9. Principal component analysis of randomly selected image patches, in which the image statistics are stationary over the patch, describes the amplitude spectrum of the patches [70] [166].

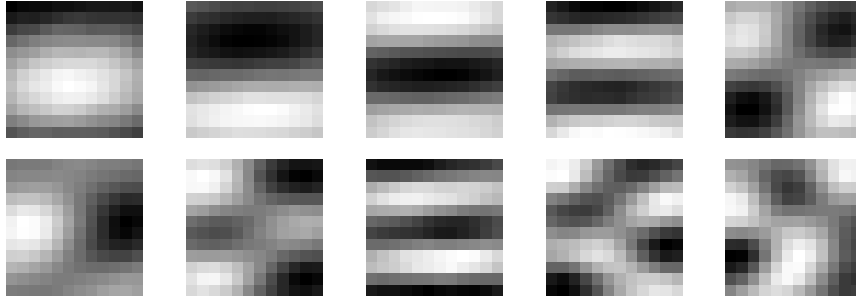


Figure 4.9: The first 10 principal components of 7750 patches extracted from random locations in the δ -images. Components are ordered left to right, top to bottom.

4.7.3 Local PCA: Fixed patches

Instead of performing PCA on patches selected from random locations in the images, we divided the images into $m \ll \frac{n}{4} 15 \times 15$ fixed regions and calculated the principal components of each region 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. The first three principal component axes of the 15×15 fixed subregions are shown in Figure 4.10.

4.7.4 Gabor wavelet representation

We next investigated classification performance with an image representation based on the outputs of local filters based on the Gabor wavelet representation. These filters closely model the receptive field properties of cells in the primary visual cortex [164] [103] [54] [50]. Such filters remove most of the variability in images due to variation in lighting and contrast. Representations of faces based on Gabor wavelets have proven successful for recognizing facial identity in images [113].

Given an image $\mathcal{I}(\vec{x})$ (where $\vec{x} = (x, y)$), the transform \mathcal{J}_i is defined as a convolution

$$\mathcal{J}_i = \int \mathcal{I}(\vec{x}) \psi_i(\vec{x} - \vec{x}') d^2 \vec{x}' \quad (4.21)$$



Figure 4.10: Local principal components of 15 x 15 patches in fixed locations of the upper face δ -images. From top to bottom: Principal components 1–3.

with a family of Gabor kernels ψ_i

$$\psi_i(\vec{x}) = \frac{\|\vec{k}_i\|^2}{\sigma^2} e^{-\frac{\|\vec{k}_i\|^2 \|\vec{x}\|^2}{2\sigma^2}} \left[e^{j\vec{k}_i \vec{x}} - e^{-\frac{\sigma^2}{2}} \right] \quad (4.22)$$

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 [113].

$$\vec{k}_i = \begin{pmatrix} f_\nu \cos \varphi_\mu \\ f_\nu \sin \varphi_\mu \end{pmatrix} \quad (4.23)$$

where

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

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, as in [113]. The Gabor filters were applied to the δ -images. 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. (See Figure 4.11 for an example.) We tested the performance of the system using $q = 1, 4, 16$ and found that $q = 16$ yielded the best generalization rate. To determine which frequency ranges contained more information for action classification, we reran the tests using subsets of high frequencies ($\nu = 0, 1, 2$), and low frequencies, ($\nu = 2, 3, 4$).

4.7.5 PCA jets

The principal components of random subimage patches are related to the Gabor representation in that they provide the amplitude spectrum of local image regions. In order to further understand the Gabor representation we developed a new representation that endowed local PCA with the multidimensionality and hierarchical properties of the Gabor wavelets (see Figure 4.12). Instead of doing PCA on random patches of fixed size, we chose five patch sizes to match the Gaussian enveloping the plane wave in each Gabor kernel. Patch sizes were chosen as $\pm 3\sigma$, yielding the following set: $[9 \times 9, 15 \times 15, 23 \times 23, 35 \times 35, \text{ and } 49 \times 49]$. 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 filter outputs were downsampled as in the Gabor representation.

4.8 Human Subjects

4.8.1 Naive subjects

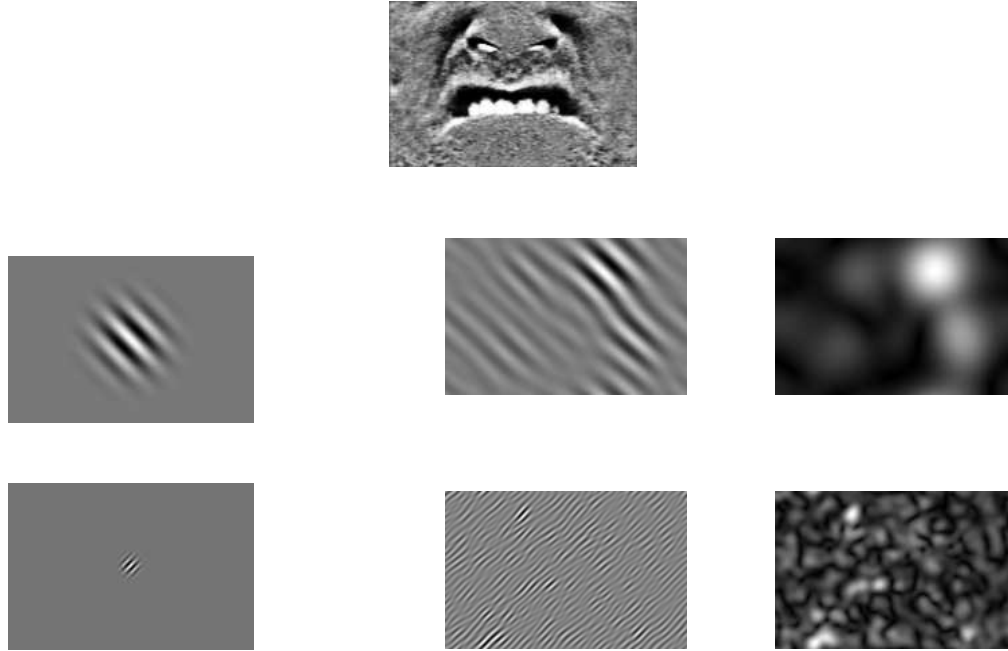


Figure 4.11: Top: Original δ -image. Bottom two rows, from left to right: Gabor kernels (low and high frequency), the imaginary part and magnitude of the filtered image.

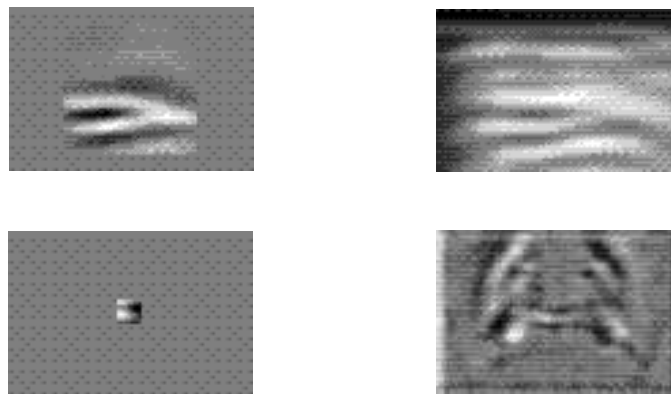


Figure 4.12: PCA Jets. Left: two kernels corresponding to low and high frequencies (patches size 49×49 and 9×9). Right: the result of the convolution with the δ -image of Figure 4.11.

A benchmark for the performance of the image classification systems was provided by the performance of naive human subjects on the same set of images. Subjects were ten adult volunteers with no prior knowledge of facial expression measurement. The task was a paper and pencil task in which images of faces were printed using a high resolution HP Laserjet 4si printer with 600 dpi. Face images were cropped and scaled identically to how they had been presented to the automated classification systems. The upper and lower face actions were tested separately. Subjects were provided with a guide sheet containing an example image of each of the six actions along with a written description of each action and a list of image cues important for detecting and discriminating the actions from [62]. Each subject was given a training session in which the facial actions were described and demonstrated, and then 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 presented in pairs, with the null image and the test image presented side by side. Subjects were instructed to compare the test image with the null 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, including low, medium, and high magnitude examples of each action. Subjects were allowed to take as much time as they needed to perform the task, which ranged from 30 minutes to one hour.

4.8.2 Expert coders

A second benchmark for the performance of the image classification systems was to compare it to that of trained FACS experts, since the objective is to ultimately replace the human experts. Subjects were four certified FACS coders. The task was a paper and pencil task as above, with the following exceptions: Expert subjects were not given a guide sheet or additional training, and performance of the experts was measured for images in which the complete face was visible, as it would normally be during FACS scoring.

The face images used in the task were cropped to display the full face from the top of the forehead to the bottom of the chin. Images were scaled to 125 x 100 pixels, with 45 pixels between the eyes, matching the spatial resolution of the upper-face images presented to the computational algorithms. One hundred and fourteen upper-face image pairs and ninety-three lower-face image pairs were presented. Subjects were instructed to take as much time as needed to complete the scoring, which ranged from 20 minutes to 1 hour and 15 minutes. Because the images were originally labeled by two expert coders with access to stop-motion video, the performance of the expert subjects provided a measure of inter-coder *agreement* for coding static images and stop-motion video.

4.9 Performance

4.9.1 Classification procedures

The output of each image analysis algorithm produced a feature vector that was used for classification. Although we use the term “feature vector” it should be emphasized that the features

were not necessarily local in nature. 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:

$$S(f^n, f^t) = \frac{\langle f^n, f^t \rangle}{\|f^n\| \cdot \|f^t\|} \in [-1, 1] \quad (4.24)$$

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 using data from all but one subject, where all of the images of that subject were reserved for testing. This procedure was repeated for each of the 20 subjects, and mean classification accuracy was calculated across all of the test cases.

The image representations were calculated using low, medium, and high magnitude facial actions. Classification performances are presented for medium magnitude facial actions, the fourth frame in each sequence. Performance results for each classifier are summarized in Table 4.1.

Classification performance was 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 results shown in Table 4.1 are for the similarity measure and classifier that gave best performance, as presented below.

4.9.2 Optic flow analysis

Flow fields were centered on the third image in the sequence, where the facial action is at medium magnitude. For the gradient-based algorithm, flow fields were calculated between the neutral image and the third image. Best performance was obtained using the cosine similarity measure and template matching.

The correlation-based flow algorithm outperformed the fast gradient-based algorithm for image classification, with 85.6% and 55.8% correct classification performances, respectively. There was a substantial difference in processing speed between the two algorithms. The correlation based method required 57 seconds to calculate a single flow field on a Dec Alpha 2100a processor compared to 0.25 seconds for the gradient based method. The addition of spatial smoothing to the correlation-based flow fields did not improve performance, and instead degraded it to 53.1%. Classification of the correlation-based optic flow algorithm is comparable to the performance of other facial expression recognition systems based on optic flow (e.g. [206] [173] See Table 4.10).

4.9.3 Holistic spatial analysis

Principal Component Analysis. Best performance with the holistic principal component representation, 79.3% correct, was obtained with the first 30 principal components, using Euclidean distance and template matching. In some previous studies, classification performance

was improved by discarding the first few principal components (c.f. [23]). For this dataset, discarding the first one and two principal components degraded performance.

Fisher’s Linear Discriminant. The dimensionality of the images was first reduced by selecting the first 30 principal components, and these were then projected down to 5 dimensions via the projection matrix, W_{fld} . Best performance of 75.7% correct was obtained with Euclidean distance and template matching. Classification performance was not improved over that obtained with the first 30 principal components by projecting into this low dimensional space. These results are consistent with other reports of poor generalization to novel subjects [41]. Good results have only been obtained with this technique when other images of the test subjects were used to calculate the projection matrix [23]. The low dimensionality appears to provide insufficient degrees of freedom for linear discrimination between classes of face images [41].

Independent component analysis. Independent component analysis performed the best of the three holistic representations. Note, however, that the independent component images in Figure 4.6 were local in nature. While the ICA algorithm analyzed the images as whole, the basis images that the algorithm settled upon were local. 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.

Local Feature Analysis. The local feature analysis representation [156] attained 81.1% correct classification using the first 155 kernels selected by the sparsification algorithm, using the cosine similarity measure and nearest neighbor classifier. LFA gave the same classification performance as global PCA.

4.9.4 Local analysis

Gaussian Kernel. A simple benchmark for the local filters consisted of a single 15 x 15 Gaussian kernel, subsampled by 0.25. The output of this basic local filter was classified at 70.3% accuracy using Euclidean distance and template matching.

PCA Random. The second set of local filters examined consisted of the principal component vectors of a set of 7000 15x15 subimage patches selected from random locations about the δ images. Images were filtered using the first p principal components, and the outputs were subsampled by a factor of 4. 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 results obtained in [153], local PCA of random patches did not outperform global PCA. The difference in performance was not statistically significant.

PCA Fixed. The representation based on PCA of fixed 15 x 15 patches gave similar classification performance to that based on PCA of random patches. 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. Performance with the local principal components of fixed patches was comparable to that with the global PCA representation.

Gabor filters. Best performance of 95.5% was obtained with the Gabor filter representation using the cosine similarity measure and nearest neighbor classifier. This finding is supported by Zhang, Yan, & Lades [212] who found that face recognition with the Gabor filter rep-

representation was superior to that with a holistic principal component based representation.

To determine which frequency ranges contain more information for action classification, we reran the classification tests using only a subset of frequencies from the Gabor filter representation. Using high frequencies only ($\nu = 0, 1, 2$) the performance of 92.8% was almost the same as for $\nu = 0, \dots, 4$, and was significantly higher ($\sim 10\%$ greater) than 83.8% yielded by the low frequencies only ($\nu = 2, 3, 4$). 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 deleterious effect on classification performance. It appears that high spatial frequencies are important for this task.

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, we developed the multiscale version of the local PCA representation, PCA jets. A multiscale local PCA representation was obtained by performing PCA on random image patches at five different scales. Sixteen principal components were retained at each scale to match the number of filters in the Gabor jets. 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 four scales, 72.1% for the two smaller scales, and 62.2% for the two larger scales. The multiscale principal component analysis (PCA jets) did not improve performance over the single scale local PCA. The multiscale property of the Gabor representation does not account for the improvement in performance obtained with this representation over local representations based on principal component analysis.

Error Analysis. Classification errors were examined for the three top performing algorithms: correlation-based flow, ICA, and Gabor wavelets. Only one image was misclassified by all 3 algorithms. AU7 was consistently misclassified as AU6 for one subject. All of the expert coders classified this image as an AU7. There was otherwise a small trend for the algorithms to make errors on the same images. The conditionally dependent error rates for each of the three algorithms, given that an image was misclassified by one of the other algorithms, were all 0.4. Due to the small N, none of these conditionally dependent error rates was significantly higher than chance.

4.10 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 [129] [206] [173] [68], holistic spatial pattern analysis using techniques based on principal components analysis [48] [153] [115], and measurements of the shapes and facial features and their spatial arrangements [115]. 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

Optic Flow				
<i>Gradient</i>	<i>Correlation</i>	<i>Smoothed</i>		
55.8% \pm 4.7%	85.6% \pm 3.3%	53.1% \pm 4.7%		

Holistic Analysis			
<i>PCA</i>	<i>FLD</i>	<i>ICA</i>	<i>LFA</i>
79.3% \pm 3.9%	75.7% \pm 4.1%	95.5% \pm 2.0%	81.1% \pm 3.7%

Local Analysis				
<i>Gaussian Kernel</i>	<i>PCA Random</i>	<i>PCA Fixed</i>	<i>PCA Jets</i>	<i>Gabor Jets</i>
70.3 \pm 4.3%	73.4% \pm 4.2%	78.3% \pm 3.9%	72.1% \pm 4.2%	95.5% \pm 2.0%

Human	
<i>Naive</i>	<i>Expert</i>
77.9% \pm 2.5%	94.1% \pm 2.1%

Table 4.1: Best performance for each classifier. FLD: Fisher’s linear discriminant. ICA: Independent component analysis. LFA: Local feature analysis.

applied them for the first time to facial expression analysis. These representations included Gabor filters [113], Linear Discriminant Analysis [23], Local Feature Analysis [156], and Independent Component Analysis [10].

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. The 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. Correlation-based optic flow performed at a level between the two, at 86%. These results compare favorably with other systems developed for emotion classification, summarized in Table 4.10.

We obtained converging evidence that local spatial filters are important for analysis of facial expressions. The two representations that gave by far the best performance were based on local filters, the Gabor representation [113] and the Independent Component representation [10]. ICA was classified as a holistic algorithm, since the analysis was performed over the images as a whole. The basis images that the algorithm produced, however, were local. 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 PCA of subimage patches performed no better than the global PCA representation (Eigenfaces). This finding differs from [153], in which a representations based on local basis functions from the principal components of local image patches outperformed global PCA. The success of the global principal component analy-

sis in this implementation could be attributable to reduced variability due to the use of difference images, or to the smaller original image size than that in [153], such that 29 principal components accounted for a greater percentage of the variability. In addition, we employed a region of interest analysis, in which images were cropped to contain half of the face, which is similar to the "Eigenfeature" approach that gave Padgett & Cottrell better performance than global-PCA.

The ICA representation performed as well as the Gabor representation, despite a two order of magnitude difference in the number of 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.

In addition to spatial locality, the ICA representation and the Gabor filter representation share the property of redundancy reduction. Relationships have been demonstrated between Gabor filters and statistical independence. Bell & Sejnowski [25] 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 pairwise independent in the presence of divisive normalization similar to the length normalization in our representation [182].

The ICA representation also captures phase information. Spatial phase contains the structural information in images that drives human recognition much more strongly than the amplitude spectrum [149] [162]. 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. The pixelwise covariances in the image set correspond to the amplitude spectrum, but not the phase spectrum, whereas the high-order statistics contain the phase information [25]. Principal component-based representations therefore contain only the amplitude spectrum of the images, whereas the independent component representation is sensitive to both amplitude and phase.

The 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 [119]. ICA allows the learning of non-orthogonal bases and allows the data to be modeled with non-Gaussian distributions [43]. 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 number of data samples is small.

We also obtained two independent 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.

Another interesting finding was that contrary to the results obtained in [23], Fisher's Linear Discriminants did not improve upon classification with PCA, 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 [23] 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 the findings of Chellappa [41] (also of H. Wechsler, personal communication) who reported that the FLD method performs well for novel images of subjects that were included in the training set, but poorly for subjects that were entirely novel. The limitation may be that FLD projects the data to a dimensionality that is *too low*. Class discriminations that are approximately linear in high dimensions may not be linear when projected down to as few as 5 dimensions.

Similarly, classification based on local feature analysis [156] also did not improve on performance with PCA representations. LFA was developed for image representation and compression, and has not been adapted for recognition. One of the authors of the method (J. Atick) obtained very high performance on the FERET face recognition test [160], but the recognition algorithm used for the test has not been disclosed. Although he has indicated that the recognition algorithm employed many of the same concepts as LFA, it should be noted that LFA was developed *after* the FERET competition.

Naive human subjects classified the facial actions at approximately the same accuracy as representations based on spatially global filters such as global-PCA, whereas the expert human subjects performance more closely matched that using spatially local filters. This supports other evidence of a shift in visual processing strategies with familiarity and expertise, from configurational processing based on external features, to local processing based on internal features [208].

The image analysis presented here made use of a 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 [153]. 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 [27], and automating this step is a realistic option. The image alignment procedure ignored out-of-plane rotations. Out-of-plane rotations could be handled by methods for estimating the frontal view of a face from a nonfrontal view [28] [201].

4.11 Conclusions

The results of this comparison provided converging evidence for the importance of possessing local filters, high spatial frequencies, and statistical independence for classifying facial actions. The relevance of high spatial frequencies has implications for motion-based facial expression analysis. Since optic flow is a noisy measure, most 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 motion in facial expression sequences is *non-rigid* and can be highly discontinuous due to the formation of wrinkles. It appears that smoothing across such boundaries is disadvantageous.

System	No. Faces	No. Classes	Method	Test Image Performance
Present Results	20	12 ^a	Gabor filters Independent components Optic flow	96% 96% 86%
Mase [129]	1	4	Optic flow	86%
Yacoob & Davis, [206]	32	7	Optic flow	87%
Kobayashi et al. [111]	30	6	Feedforward neural network taking pixel graylevels as input	90%
Rosenblum, et al. [173]	32	2	Radial basis functions; optic flow	88%
Padgett & Cottrell [153]	12	6 ^b	PCA of static graylevel images	86%
Essa & Pentland [68]	8	5	Parameterized face model; optic flow	98%
Lanitis et al. [115]	30	7	Adaptive local features; PCA of static graylevel images	74%
Cohn et al. [42]	30	15 ^a	Feature tracking based on optic flow	86%

Table 4.2: Summary of automated facial expression analysis systems. These models were tested on different data sets with differing levels of difficulty. Classification categories were feigned expressions of emotion, except where otherwise indicated. ^aClassified facial actions ^bClassified images from Pictures of Facial Affect (Ekman & Friesen, 1976) in which trained actors performed the muscle contractions empirically associated with certain emotional states.

The majority of the approaches to facial expression recognition by computer have focused exclusively on analysis of facial motion. It should be noted that while human subjects *can* recognize facial expressions from motion signals alone [15], recognition rates are just above chance, and much lower than for static graylevel images. In this comparison, we found that best performance was obtained with representations based on surface graylevels. A future direction of this work is to combine the best motion classifiers with the best classifiers for static graylevel images. Perhaps combining motion and graylevel information will ultimately provide the best facial expression recognition performance, as it does for human subjects [202].

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 Linda Camras, Wil Irwin, Irene McNee, Harriet Oster, and Erica Rosenberg for their time and assistance with this project. We thank Paul Viola and Jan Larsen for contributions to algorithm development, Wil Irwin and Beatrice Golomb for contributions to project initiation, Claudia Hilburn for image collection, and Laurenz Wiskott for valuable discussions on earlier drafts of this paper.

This chapter, in full, is a reprint of material that was submitted for publication in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. Submitted March 25, 1988. The dissertation author was secondary researcher, primary writer, and project supervisor.

Chapter 5

Learning Viewpoint Invariant Face Representations from Visual Experience in an Attractor Network

5.1 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 [72] 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 [12]. 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.

5.2 Introduction

Cells in the primate inferior temporal lobe have been reported that respond selectively to faces despite substantial changes in viewpoint [158, 87]. A small proportion of cells gave responses that were invariant to angle of view, whereas other cells that have been classed as view-

Copyright 1998, Institute of Physics Publishing, Ltd. Reprinted with permission from *Network: Computation in Neural Systems* 9(3) 399-417, 1998.

point *dependent* had tuning curves that were quite broad. Perrett et al. [158] 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 $\pm 40^\circ$. 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. [32]). 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 [190].

Temporal association may be an important factor in the development of pose invariant responses in the inferior temporal lobe of primates [171]. 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 [135]. Macaques were presented a fixed sequence of 97 fractal patterns for 2 weeks. After training, the patterns were presented in random order. Figure 5.1 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 [72]. This mechanism learns viewpoint-tolerant representations when different views of an object are presented in temporal continuity [72, 205, 167, 150, 204]. Földiák [72] 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 [171]. Most such models used idealized input representations. These learning mechanisms have recently been shown to learn transformation invariant of responses to complex inputs such as im-

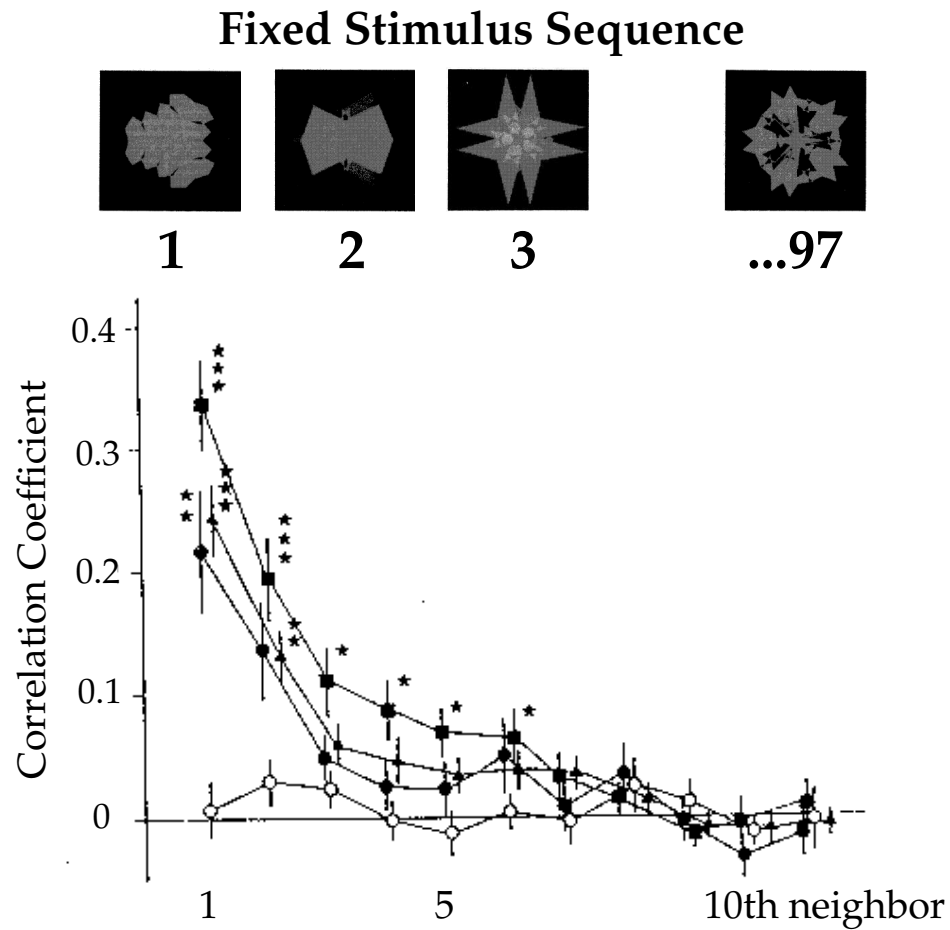


Figure 5.1: Evidence of temporal associations in IT. Top: Samples of the 97 fractal pattern stimuli in the fixed training sequence. Bottom: Autocorrelograms on the sustained firing rates of AIT cells along the serial position number of the stimuli. Abscissa is the relative position of the patterns in 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 *Nature*, copyright 1988, MacMillan Magazines, Ltd.

ages of faces [12, 13, 204, 19]. 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 [18, 186].

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 [167, 170]. A spatio-temporal window for Hebbian learning could also be produced by the release of a chemical signal following activity such as nitric oxide [136]. Recurrent excitatory connections within a cortical area and reciprocal connections between cortical regions [150] 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 [167]. 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 [174] is extended to incorporate an activity trace on the output unit activities [72]. 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 [4, 74, 95]. Weinshall and Edelman [205] 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 [97] 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 [80].

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.

5.3 Simulation

Stimuli for these simulations consisted of 100 images of faces undergoing a change in pose, from David Beymer [27] (see Figure 5.2). 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 [27]. 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.

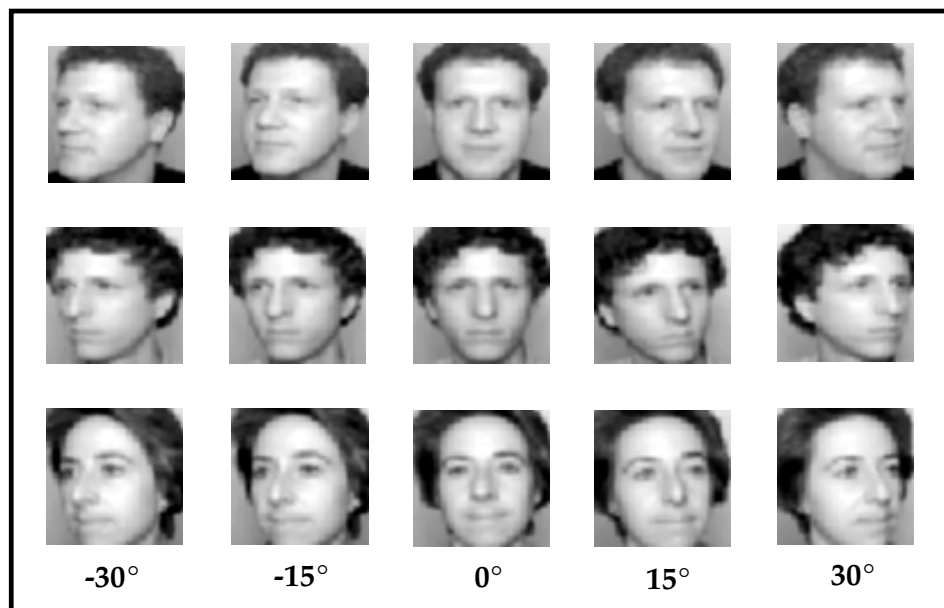


Figure 5.2: Sample of the 100 images used in the simulation. Image set provided by David Beymer (1994).

5.3.1 Model architecture

Images were presented to the model in sequential order as the subject changed pose from left to right (Figure 5.3). The first layer of processing consisted of an oriented energy model related to the output of V1 complex cells [50, 113]. 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 [90]. 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.

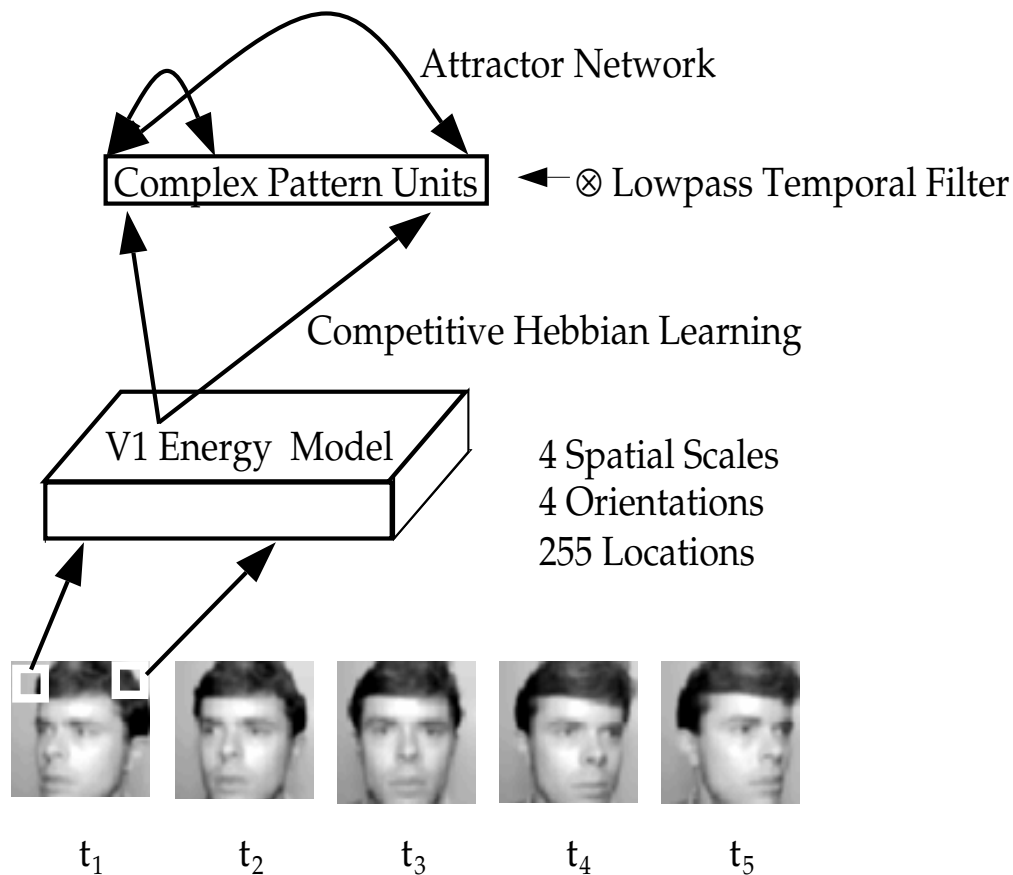


Figure 5.3: Model architecture.

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.

5.3.2 Competitive Hebbian learning of temporal relationships

The learning rule for the feedforward connections of the model was an extension of the Competitive Learning Algorithm [174, 81]. The output unit activities were passed through a lowpass temporal filter [12]. 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:

$$\overline{y}_j^{(t)} = (1 - \lambda)y_j^t + \lambda\overline{y}_j^{(t-1)} \quad (5.1)$$

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

$$V_j = \begin{cases} 1 & \text{if } j = \max_j [\overline{y}_j^{(t)}] \\ \frac{\alpha}{N} & \text{otherwise} \end{cases} \quad (5.2)$$

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 [174]. The feedforward connections were updated according to the following learning rule:

$$\Delta w_{ij} = \alpha V_j \left(\frac{x_{iu}}{\sum_k x_{ku}} - w_{ij} \right) \quad (5.3)$$

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

$$\Delta b_j = \beta \left(\frac{P}{n} - c_j \right) \quad (5.4)$$

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 activity dependent modification of receptive field size of cortical neurons (eg. [102, 104]). An alternative way to normalize responses is through multiplicative scaling of the synaptic weights [198].

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 [98]. When the summed lengths of the optic flow vectors for sequential image pairs exceeded a threshold of $\gamma = 25$, \bar{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 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 components 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 [146, 177]. The low-pass temporal filter on output unit activities in Equation 5.1 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.

5.3.3 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 5.1) 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:

$$W_{ij} = \frac{1}{N} \sum_{t=1}^P (y_i^t - y^0)(y_j^t - y^0) \quad (5.5)$$

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 $\bar{y}_i^{(t)}$ defined in Equation 5.1, we obtain

$$W_{ij} = \frac{1}{N} \sum_{t=1}^P \left((1 - \lambda)y_i^t + \lambda\bar{y}_i^{(t-1)} - y^0 \right) \left((1 - \lambda)y_j^t + \lambda\bar{y}_j^{(t-1)} - y^0 \right) \quad (5.6)$$

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

¹This 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. [203]). 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.

$$\begin{aligned}
W_{ij} = & \frac{1}{N} \sum_{t=1}^P ((1 - \lambda)^2 (y_i^t - y^0)(y_j^t - y^0) \\
& + \lambda(1 - \lambda) \left[(y_i^t - y^0)(\bar{y}_j^{(t-1)} - y^0) + (\bar{y}_i^{(t-1)} - y^0)(y_j^t - y^0) \right] \\
& + \lambda^2 \left[(\bar{y}_i^{(t-1)} - y^0)(\bar{y}_j^{(t-1)} - y^0) \right]) \quad (5.7)
\end{aligned}$$

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 [80]. 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 [80]:

$$V_i(t + \delta t) = \phi \left[\sum W_{ij} V_j(t) - \theta \right] \quad (5.8)$$

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 [80] is presented in Equation 5.9 for comparison. The Griniasty et. al. learning rule associates first neighbors in the pattern sequence, whereas the learning rule in 5.7 has a longer memory. The weights in 5.9 are a function of the *discrete* activities at t and $t - 1$, whereas the weights in 5.7 are a function of the current input and the activity *history* at time $t - 1$.

$$W_{ij} = \frac{1}{N} \sum_{t=1}^P (y_i^t - y^0)(y_j^t - y^0) + a \left[(y_i^{t+1} - y^0)(y_j^t - y^0) + (y_i^t - y^0)(y_j^{t+1} - y^0) \right] \quad (5.9)$$

The weight structure and fixed points of an attractor network trained with Equation 5.7 are illustrated in Figures 5.4 and 5.5 using an idealized data set in order to facilitate visualization. The fixed points for the real face data will be illustrated later, in Section 5.3.4. The idealized data set contained 25 input patterns, where each pattern was coded by activity in a single bit (Figure 5.4, Top). The patterns represented 5 individuals with 5 views each (a - e). The middle graph in Figure 5.4 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 net, 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 5.5 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 [140].

5.3.4 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 5.1-5.3, 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.

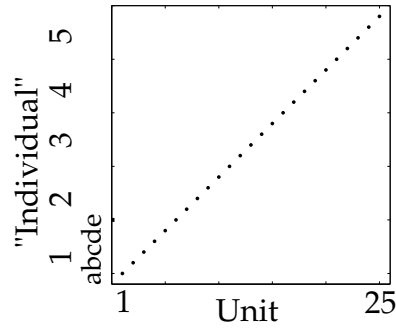
After training the feedforward connections, the representation of each face was a sparse representation consisting of the two active output units out of the total of 70 complex pattern units. “Pose tuning” of the feedforward system was assessed by comparing correlations in the network outputs for different 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° changes in pose, three 30° 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° changes in pose, 30° changes in pose, and so forth.

Figure 5.6 (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

²The half-life, h , of the temporal filter is related to λ by $\lambda^h = 0.5$ [186]. For $\lambda = 0.5$, the activity at time t is reduced by 50% after one time step.

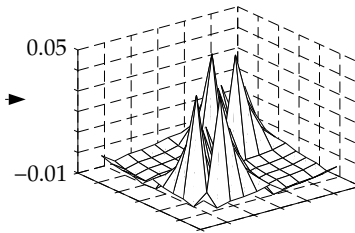
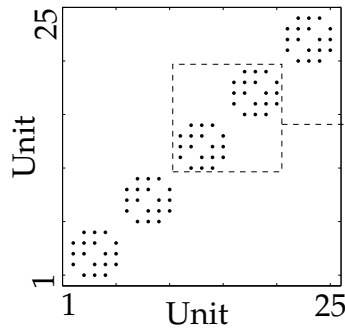
Input Patterns



25 patterns
5 "Individuals"
5 "views"
25 units
Coded by single bit

Weight Matrix

$$\lambda = 0.5$$



Activity States

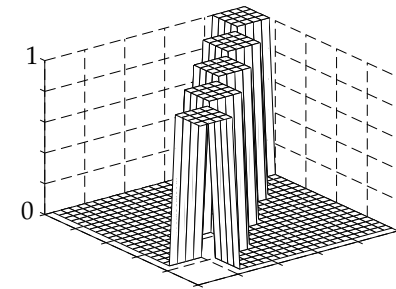
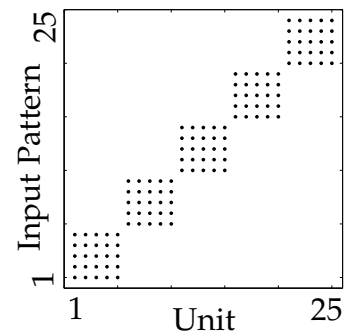
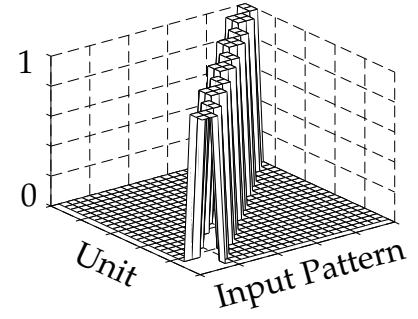
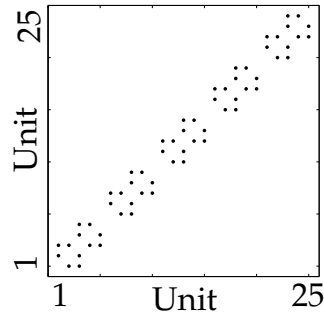


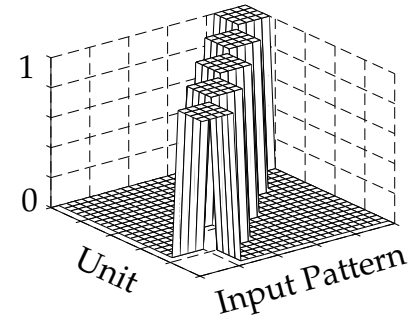
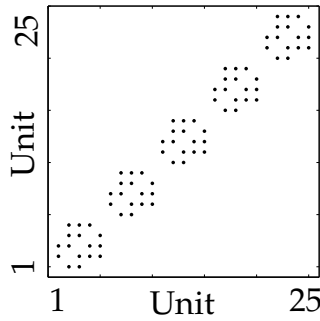
Figure 5.4: Demonstration of attractor network with idealized data. Top: Idealized data set. The patterns consist of 5 "individuals" (1,2,3,4,5) with five "views" each (a,b,c,d,e), and are each coded by activity in 1 of the 25 units. Center: The weight matrix obtained with equation 3. Dots show the locations 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.

Temporal Filter**Weight Matrix****Activity States**

$$\lambda = 0.25$$



$$\lambda = 0.4$$



$$\lambda = 0.6$$

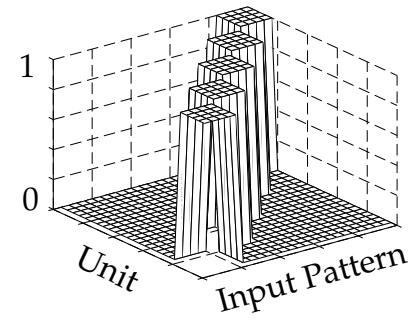
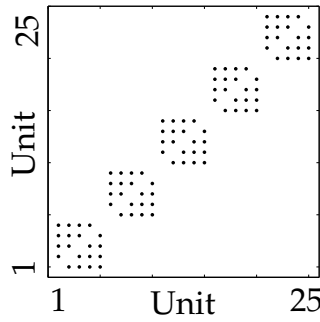


Figure 5.5: Weight matrix (left) and fixed points (right) for three values of the temporal filter, λ . Dots show locations of positive weights. Unit activities are plotted for each of the 25 input patterns of the simplified data.

λ_2	λ_1	
	0	0.5
0	.70	.90
0.5	.84	.98

Table 5.1: 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).

against the proportion of false alarms (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 5.6 (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 produced a similar pattern of results, which are presented in the bottom of Figure 5.6.

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 5.7. 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 5.7 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 5.7. 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 5.1. 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 5.8 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

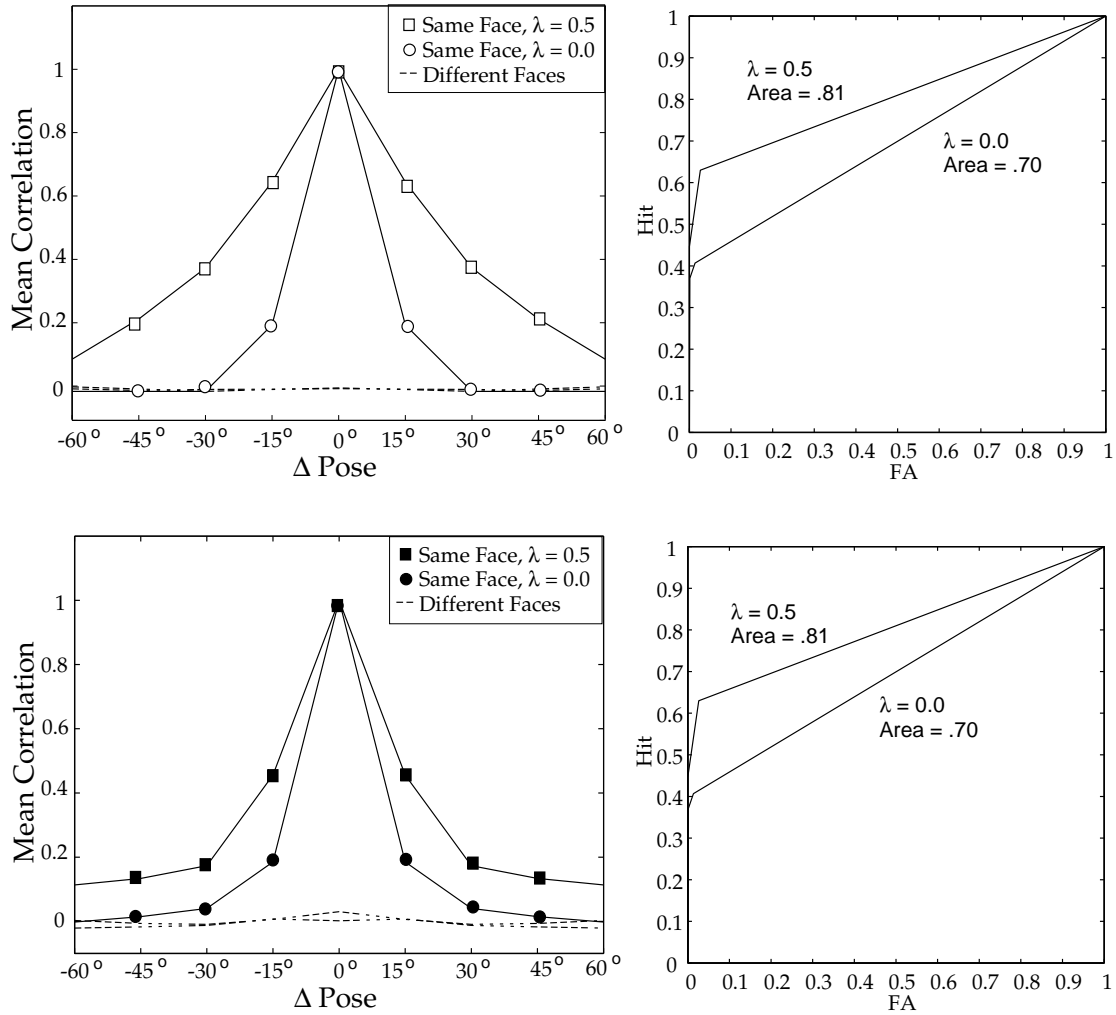


Figure 5.6: Pose tuning and ROC curves of the feedforward system for training images (top) and test images (bottom). Left: Mean correlations of the feedforward system outputs for pairs of face images are presented by change in pose. Correlations across different views of the same face (—) are compared 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).

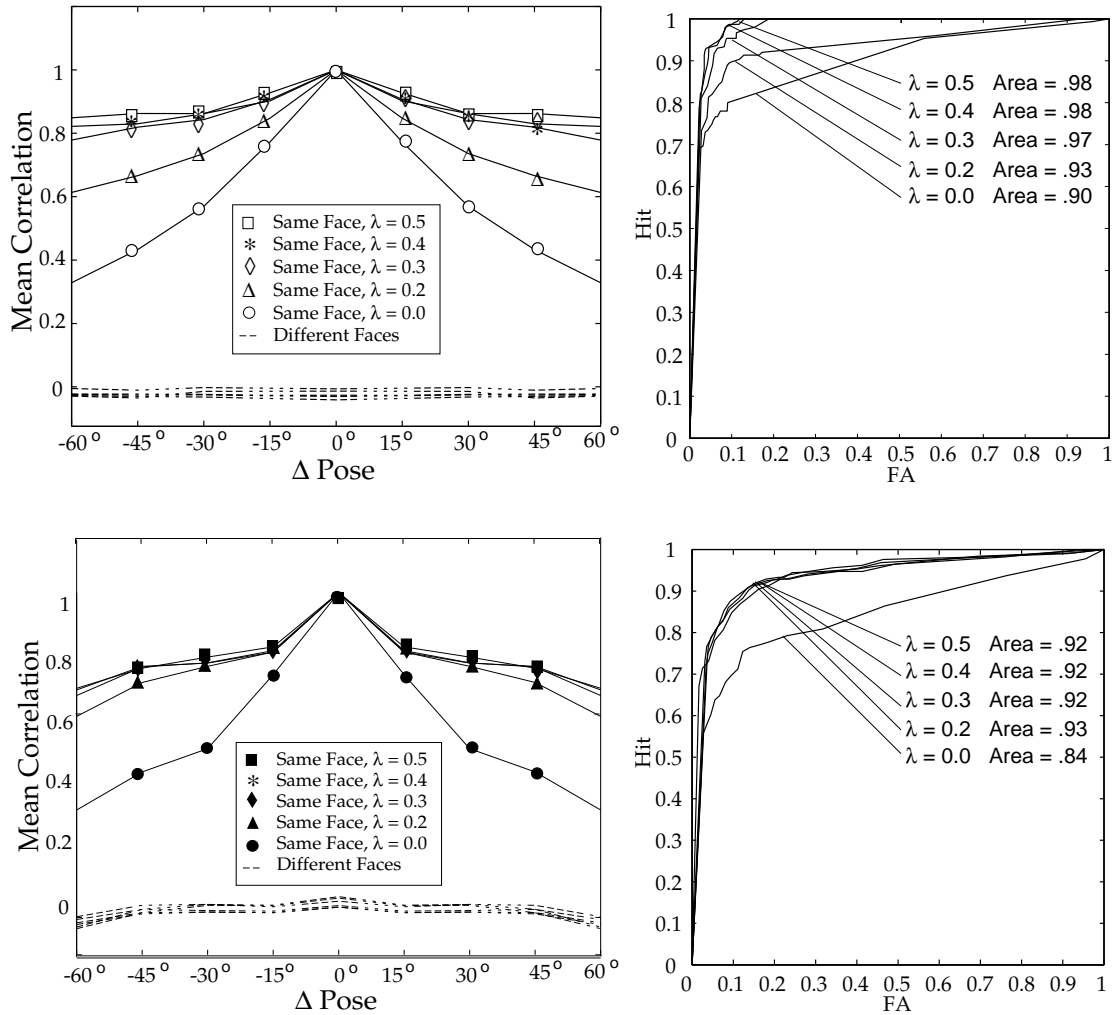


Figure 5.7: Pose tuning and ROC curves of the attractor network for training images (top) and test images (bottom). Left: Mean correlations in sustained activity patterns in the attractor network for pairs of face images are presented by change in pose. Correlations across different views of the same 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).

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 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 [205] for idealized wire-framed objects, units that were active for one view of a face in the input to the 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 presented elsewhere [140] (see also [76, 195]).

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 $(y_0 |\ln(y_0)|)^{-1}$ [76, 195]. 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 [195, 140]:

$$\frac{F}{N} \leq \frac{0.2}{s^2 y_0 \ln\left(\frac{1}{s y_0}\right)} \quad (5.10)$$

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 [75] and area TE, a posterior IT region which projects to the anterior IT regions where transformation invariant responses can be found [192]. 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 faces on average in the attractor network, compared to 3 faces for its

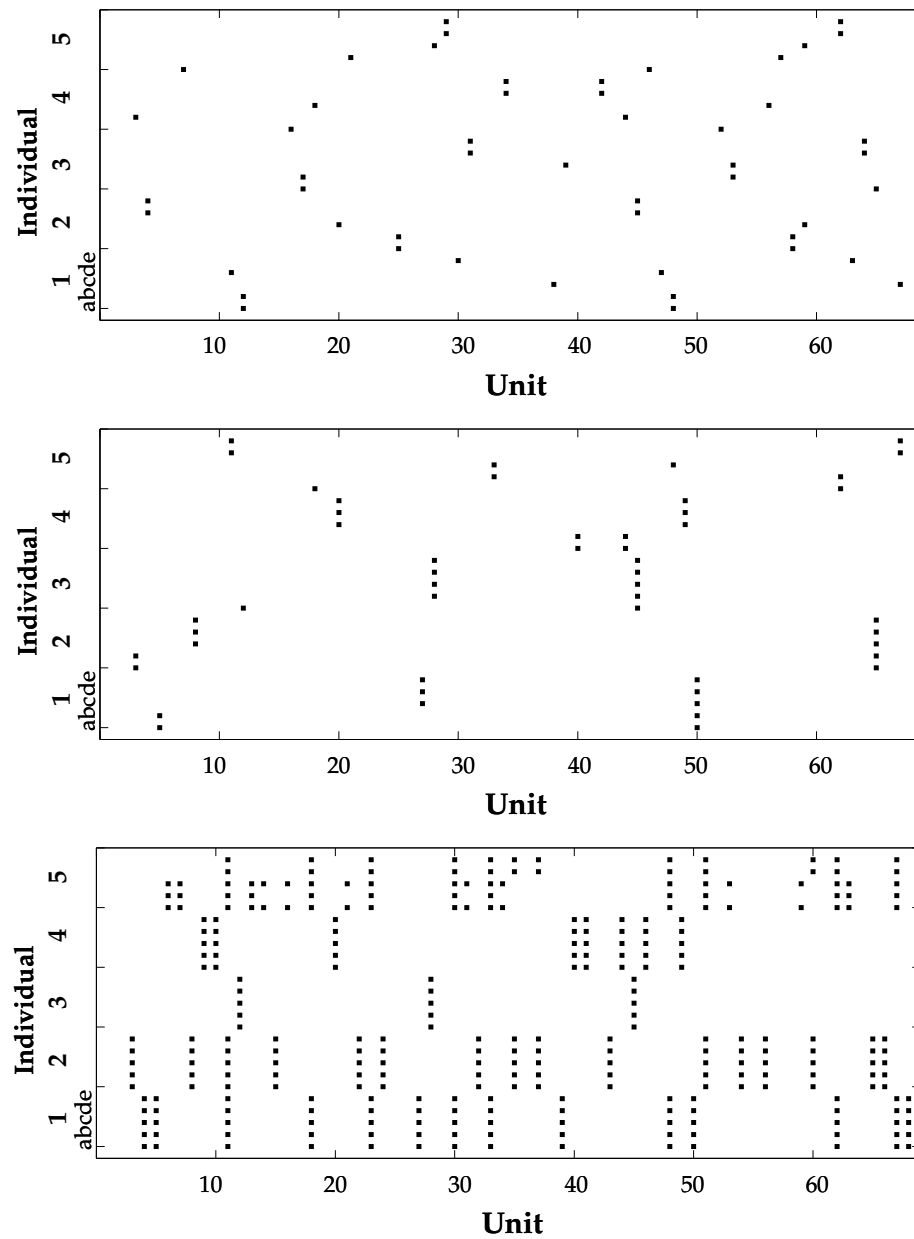


Figure 5.8: Coding of real face image data. Top: Coding of 5 faces in network layer 2 following training of the feedforward connections only, with no temporal lowpass filter ($\lambda = 0$.) The vertical axis is the input image, with the five poses of each individual labeled a,b,c,d,e. The two active 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 $\lambda = 0.5$.

F	P	Attractor Network			Eigenfaces
		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 5.2: 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 $\frac{F}{N} < 0.14$

input. The coding levels in the attractor network were consistent with the sparse-distributed face coding reported in IT [209, 1].

We evaluated face recognition performance of the attractor network using a nearest neighbor classifier on the sustained activity patterns at several loading levels. Table 5.2 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 on the load parameter, $\frac{F}{N}$. Performance was quite good when $\frac{F}{N} \ll 0.14$, 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 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 5.4 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%.

5.4 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 [87, 158, 125]. 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 [167], 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 [114]. Recognition of novel views of unfamiliar faces was superior when the faces were presented in continuous motion during learning [161]. Stone [187] 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 [204] trained a hierarchical feedforward system using Hebbian learning and the temporal activity trace of Equation 5.1. Their system successfully learned translation invariant representations of seven faces, and rotation invariant representations of three faces. Parga and Rolls [140] 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 components analysis of the Gabor filter representation. It has been shown that competitive Hebbian learning finds the principal components of the input data [146, 177]. 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 [199] or nonlinear combination of stored 2-D views [163, 35]. 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 [29, 33].

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 [205]. 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 as evidence for mental rotation of an internal 3-D model [181].

In example-based models of recognition such as radial basis functions [163], 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 [125] 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 $\pm 10^\circ$ 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.

This chapter, in full, is a reprint of material published in *Network: Computation in Neural Systems* 9(3), 399-417, 1998, Bartlett, M.S., & Sejnowski, T.J. The dissertation author was primary investigator and author of this paper. Figure 5.1 was reprinted with permission from *Nature* 335, copyright 1988, MacMillan Magazines Ltd.

Chapter 6

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 [8]. 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" [197] "Holons" [48], and "Local Feature Analysis" [156] 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 [24]. This algorithm is an unsupervised learning rule derived from the principle of optimal information transfer through sigmoidal neurons [116, 6], and information maximization [124]. 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 2 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 recognizing images collected within the same session as well. We have demonstrated elsewhere that ICA representations can outperform PCA representations for recognizing faces across changes in pose, and changes in lighting [13].

Chapters 3 and 4 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 [62]. Chapter 3 showed that a PCA representation gave better recognition performance than a set of hand-crafted feature measurements. The results also suggest that hand-crafted features plus principal component representations may be superior to either one alone, since their performances may be uncorrelated.

Chapter 4 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 [25] 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 [182] 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 4.

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 5 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 [72] 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-crafted 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.

As discussed in Chapters 1 and 2, the ICA algorithm produces sparse outputs [25]. Due to the advantages of sparse codes for associative memory [16], the ICA factorial representation would constitute a good input representation for the attractor network model of Chapter 5. Preliminary explorations demonstrated the success of this implementation for learning pose invariant representations of faces [11].

The model of Chapter 5 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 (James Stone, personal communication). 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 [188].

Another area for exploration is methods for extracting fewer sources than mixtures for the independent component representations presented in Chapter 2. In Chapter 2, the number of sources was controlled by reducing the dimensionality of the data through principal component analysis prior to performing ICA. There are two limitations to this approach (James Stone, personal communication). 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 2 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 are presently under development for separating sources on projection planes without throwing away subspaces of the input data [165].

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 [24]. 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 [117], 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 [84]. 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 [118].

A second approach to independent component analysis involves building a generative model of the data using maximum likelihood methods [126]. 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 be 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 [120]. 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

- [1] L. Abbott, R. E., and M. Tovee. Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6(3):498–505, 1996.
- [2] R. Adolphs, H. Damasio, D. Tranel, and A. Damasio. Cortical systems for the recognition of emotion in facial expressions. *Journal of Neuroscience*, 16(23):7678–7687, 1996.
- [3] R. Adolphs, D. Tranel, H. Damasio, and A. Damasio. Fear and the human amygdala. *Journal of Neuroscience*, 15(9):5879–5891, 1995.
- [4] D. Amit. The hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences*, 18:617–657, 1995.
- [5] A. Antonini and M. Stryker. Development of individual geniculocortical arbors in cat striate cortex and effects of binocular impulse blockade. *Journal of Neuroscience*, 13(8):3549–73, 1993.
- [6] J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.
- [7] J. Atick and A. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992.
- [8] H. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.
- [9] H. Barlow. What is the computational goal of the neocortex? In C. Koch, editor, *Large scale neuronal theories of the brain*, pages 1–22. MIT Press, Cambridge, MA, 1994.
- [10] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representations for face recognition. In T. Rogowitz, B. & Pappas, editor, *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, volume 3299, San Jose, CA, January 1998. SPIE Press.
- [11] M. Bartlett and T. Sejnowski. learning viewpoint invariant representations of faces in an attractor network. In G. Cottrell, editor, *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, page 730, July 1996.

- [12] M. Bartlett and T. Sejnowski. Unsupervised learning of invariant representations of faces through temporal association. In J. Bower, editor, *Computational Neuroscience: Trends in Research; Int. Rev. Neurobio. Suppl. 1*, pages 317–322, San Diego, CA, 1996. Academic Press.
- [13] M. Bartlett and T. Sejnowski. Viewpoint invariant face recognition using independent component analysis and attractor networks. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 817–823, Cambridge, MA, 1997. MIT Press.
- [14] M. Bartlett, P. Viola, T. Sejnowski, J. Larsen, J. Hager, and P. Ekman. Classifying facial action. In D. Touretski, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 823–829. Morgan Kaufmann, San Mateo, CA, 1996.
- [15] J. Bassili. 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, 1979.
- [16] E. Baum, J. Moody, and F. Wilczek. Internal representations for associative memory. *Biological Cybernetics*, 59:217–228, 1988.
- [17] S. Becker. *An Information-theoretic Unsupervised Learning Algorithm for Neural Networks*. PhD thesis, University of Toronto, 1992.
- [18] S. Becker. Learning to categorize objects using temporal coherence. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 361–368, San Mateo, CA, 1993. Morgan Kaufmann.
- [19] S. Becker. Implicit learning in 3d object recognition: The importance of temporal context. *Neural Computation*, in press.
- [20] S. Becker and G. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 335(6356):161–3, 1992.
- [21] S. Becker and G. Hinton. Learning mixture models of spatial coherence. *Neural Computation*, 5:267–277, 1993.
- [22] S. Becker and M. Plumbley. Unsupervised neural network learning procedures for feature extraction and classification. *Journal of Applied Intelligence*, 6:1–21, 1996.
- [23] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
- [24] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

- [25] A. Bell and T. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.
- [26] G. Berns, P. Dayan, and T. Sejnowski. 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, 1993.
- [27] D. Beymer. 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, 21-23 June 1994. IEEE Comput. Soc. Press, Los Alamitos, CA.
- [28] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. AI Memo 1431, Massachusetts Institute of Technology, 1993.
- [29] I. Biederman. Neural and psychophysical analysis of object and face recognition. In H. . Wechsler, P. Phillips, V. Bruce, F. Fogelman-Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag, in press.
- [30] C. Blakemore. Sensitive and vulnerable periods in the development of the visual system. *Ciba Foundation Symposium*, 156:129–47, 1991.
- [31] H. Breiter, N. Etcoff, P. Whalen, W. Kennedy, S. Rauch, R. Buckner, M. Strauss, S. Hyman, and B. Rosen. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17(5):875–87, 1996.
- [32] V. Bruce. *Recognising Faces*. Lawrence Erlbaum Assoc., London, 1988.
- [33] V. Bruce. Human face perception and identification. In H. . Wechsler, P. Phillips, V. Bruce, F. Fogelman-Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag, in press.
- [34] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052, 1993.
- [35] H. Bulthoff, S. Edelman, and M. Tarr. How are three-dimensional objects represented in the brain. *Cerebral Cortex*, 3:247–260, 1995.
- [36] J. Cacioppo, L. Tassinari, and A. Friedlund. The skeletomotor system. In *Principles of Psychophysiology: Physical, Social, and Inferential Elements*, pages 325–384. Cambridge University Press, New York, 1990.
- [37] E. Callaway and L. Katz. 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, 1991.
- [38] L. Camras. Facial expressions used by children in conflict situations. *Child Development*, 48:1431–35, 1977.

- [39] L. Camras, H. Oster, J. Campos, K. Miyake, and D. Bradshaw. Japanese and american infants responses to arm restraint. *Developmental Psychology*, 28:578–583, 1992.
- [40] G. Carmignoto and S. Vicini. Activity dependent decrease in nmda receptor responses during development of the visual cortex. *Science*, 1992.
- [41] R. Chellappa. Discriminant analysis for face recognition. In H. Wechsler, P. Phillips, V. Bruce, F. Fogelman-Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications. NATO ASI Series F*. Springer-Verlag, in press.
- [42] J. Cohn, A. Zlochower, J. Lien, Y.-T. Wu, and T. Kanade. Automated face coding: A computer-vision based method of facial expression analysis. *Psychophysiology*, in press.
- [43] P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.
- [44] M. Constantine-Paton, H. Cline, and E. DeBinski. Patterend activity, synaptic convergence, and the nmda receptor in developing visual pathways. *Annual Review of Neuroscience*, 13:129–154, 1990.
- [45] E. Cooper. Unpublished research, as cited by I. Biederman, (in press), in H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman-Soulie, T. Huang, (Eds.), *Face Recognition: From Theory to Applications*. Springer-Verlag.
- [46] G. Cottrell. Extracting features from faces using compression networks: Face, identity, emotion, and gender recognition using holons. Connectionist Models Summer School, 1990.
- [47] G. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference*, pages 322–325, Dordrecht, 1990. Kluwer.
- [48] G. Cottrell and J. . Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufmann.
- [49] K. Craig, S. Hyde, and C. Patrick. Genuine, suppressed, and faked facial behavior during exacerbation of chronic low back pain. *Pain*, 46:161–172, 1991.
- [50] J. Daugman. Complete discrete 2d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:1169–1179, 1988.
- [51] R. Davidson, P. Ekman, C. Saron, J. Senulis, and W. Friesen. Emotional expression and brain physiology i. approach/withdrawal and cerebral asymmetry. *Journal of Personality and Social Psychology*, 58:330–341, 1990.

- [52] P. Dayan, G. Hinton, R. Neal, and R. Zemel. The helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.
- [53] E. Debinski, H. Cline, and M. Constantine-Paton. Activity-dependent tuning and the nmda receptor. *Journal of Neurobiology*, 21(1):18–32, 1990.
- [54] R. DeValois and K. DeValois. *Spatial Vision*. Oxford Press, 1988.
- [55] P. Ekman. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology*, pages 169–248. Cambridge Univ Press, New York, 1979.
- [56] P. Ekman. *Emotion in the Human Face, Second Edition*. Cambridge University Press, New York, 1982.
- [57] P. Ekman. Methods for measuring facial action. In K. Scherer and P. Ekman, editors, *Handbook of Methods in Nonverbal Behavior Research*, pages 45–135. Cambridge University Press, New York, 1982.
- [58] P. Ekman. Expression and the nature of emotion. In K. Scherer and P. Ekman, editors, *Approaches to Emotion*, pages 319–343. Lawrence Erlbaum, Hillsdale, NJ, 1984.
- [59] P. Ekman. The argument and evidence about universals in facial expressions of emotion. In D. Raskin, editor, *Psychological methods in criminal investigation and evidence*, pages 297–332. Springer Publishing Co, Inc., New York, 1989.
- [60] P. Ekman. Facial expression of emotion. *American Psychologist*, 48:384–392, 1993.
- [61] P. Ekman and W. Friesen. *Unmasking the Face; A Guide to Recognizing Emotions from Facial Clues*. Prentice-Hal, NJ, 1975.
- [62] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.
- [63] P. Ekman, W. Friesen, and M. O’Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 54:414 – 420, 1988.
- [64] P. Ekman, T. Huang, T. Sejnowski, and J. Hager. Final report to NSF of the planning workshop on facial expression understanding, 1992. Available from UCSF, HIL-0984, San Francisco, CA 94143.
- [65] P. Ekman, R. Levenson, and W. Friesen. Autonomic nervous system activity distinguishes between emotions. *Science*, 221:1208–1210, 1983.
- [66] P. Ekman and H. Oster. Facial expressions of emotion. *Annual Review of Psychology*, 30:527–554, 1979.

- [67] P. Ekman and E. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York, 1997.
- [68] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–63, 1997.
- [69] D. Field. 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, 1987.
- [70] D. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.
- [71] R. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.
- [72] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.
- [73] K. Fox, N. Daw, H. Sato, and D. Czepita. The effect of visual experience on the development of nmda receptor synaptic transmission in kitten visual cortex. *Journal of Neuroscience*, 13:155–69, 1992.
- [74] W. Freeman. Characterization of state transitions in spatially distributed, chaotic, non-linear, dynamical systems in cerebral cortex. *Integrative Physiological and Behavioral Science*, 29(3):294–306, 1994.
- [75] J. Gallant, C. Connor, and D. Van Essen. Responses of visual cortical neurons in a monkey freely viewing natural scenes. In *Society for Neuroscience Abstracts*, volume 20, page 838, 1994.
- [76] E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Math. Gen.*, 21:257–270, 1988.
- [77] J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum, New Jersey, 1986.
- [78] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In R. Lippman, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, pages 572–577. Morgan-Kaufmann, San Mateo, CA, 1991.
- [79] M. Gray, J. Movellan, and T. Sejnowski. 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, 1997.

- [80] M. Griniasty, M. Tsodyks, and D. Amit. Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 5:1–17, 1993.
- [81] S. Grossberg. Adaptive pattern classification and universal recoding: Part 1. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134, 1976.
- [82] Q. Gu and W. Singer. Effects of intracortical infusion of anticholinergic drugs on neuronal plasticity in kitten striate cortex. *European Journal of Neuroscience*, 5(5):475–85, 1993.
- [83] J. Hager and P. Ekman. The essential behavioral science of the face and gesture that computer scientists need to know. In M. Bichsel, 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, 1995.
- [84] P. Hallinan. *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions*. PhD thesis, Harvard University, 1995.
- [85] P. Hancock, A. Burton, and V. Bruce. Face processing: human perception and principal components analysis. *Memory and Cognition*, 24:26–40, 1996.
- [86] W. Harris and C. Holt. Early events in the embryogenesis of the vertebrate visual system: Cellular determination and path finding. *Annual Review of Neuroscience*, 13:155–169, 1990.
- [87] M. Hasselmo, E. Rolls, G. Baylis, and V. Nalwa. 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, 1989.
- [88] S. Haykin. *Neural Networks: A Comprehensive Foundation*. MacMillan, New Jersey, 1994.
- [89] D. Hebb. *The organization of Behavior*. Wiley, New York, 1949.
- [90] D. Heeger. Nonlinear model of neural responses in cat visual cortex. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*, pages 119–133. MIT Press, Cambridge, MA, 1991.
- [91] D. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197, 1992.
- [92] M. Heller and V. Haynal. The faces of suicidal depression (translation). les visages de la depression de suicide. *Kahiers Psychiatriques Genevois (Medecine et Hygiene Editors)*, 16:107–117, 1994.
- [93] W. Himer, F. Schneider, G. Kost, and H. Heimann. Computer-based analysis of facial action: A new approach. *Journal of Psychophysiology*, 5(2):189–195, 1991.

- [94] G. Hinton, P. Dayan, B. Frey, and R. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–61, 1995.
- [95] G. Hinton and T. Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–75, 1991.
- [96] G. Hinton and R. Zemel. Autoencoders, minimum description length, and helmholtz free energy. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 3–10, San Francisco, CA, 1994. Morgan Kaufmann.
- [97] J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, 79:2554–2558, 1982.
- [98] B. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185 – 203, 1981.
- [99] T. Huang. Face recognition using 3-d models. In H. Wechsler, P. Phillips, V. Bruce, S. Fogelman-Soulie, , and T. Huang, editors, *Face Recognition from Theory to Applications; NATO ASI Series F*. Springer-Verlag, in press.
- [100] D. Hubel, T. Wiesel, and S. LeVay. Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical transactions of the Royal Society of London (Biol.)*, 278:377–409, 1977.
- [101] C. Izard. The maximally discriminative facial movement coding system (max). Available from Instructional Resource Center, University of Delaware, Newark, Delaware., 1979.
- [102] W. Jenkins, M. Merzenich, and G. Recanzone. Neocortical representational dynamics in adult primates: implications for neuropsychology. *Neuropsychologia*, 28(6):573–84, 1990.
- [103] J. Jones and L. Palmer. An evaluation of the two dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58:1233–1258, 1987.
- [104] J. Kaas. Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscience*, 14:137–67, 1991.
- [105] S. Kaiser and T. Wherle. Automated coding of facial behavior in human-computer interactions with facs. *Journal of Nonverbal Behavior*, 16(2):65–140, 1992.
- [106] T. Kanade. *Computer recognition of human faces*. Birkhauser Verlag, Basel and Stuttgart, 1977.
- [107] S. Kanfer. *Serious business : the art and commerce of animation in America from Betty Boop to Toy story*. Scribner, New York, 1997.
- [108] D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68:441–454, 1995.

- [109] S. Kinomura, R. Kawashima, K. Yamada, S. Ono, M. Itoh, S. Yoshioka, T. Yamaguchi, H. Matsui, H. Miyazawa, H. Itoh, and et al. Functional anatomy of taste perception in the human brain studied with positron emission tomography. *Brain Research*, 659(1-2):263–6, 1994.
- [110] D. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge University Press, New York, 1996.
- [111] H. Kobayashi, K. Tange, and F. Hara. Real-time recognition of six basic facial expressions. In *Proceedings 4th IEEE International Workshop on Robot and Human Communication*, pages 179–86, Tokyo, Japan, 5-7 July 1995.
- [112] T. Kohonen, E. Oja, and P. Lehtio. Storage and processing of information in distributed associative memory systems. In G. Hinton and J. Anderson, editors, *Parallel Models of Associative Memory*, pages 49–81. Erlbaum, Hillsdale, 1981.
- [113] M. Lades, J. Vorbrüggen, J. Buhmann, J. Lange, W. Konen, C. von der Malsburg, and R. Würtz. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.
- [114] K. Lander and V. Bruce. 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., July 1997.
- [115] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.
- [116] S. Laughlin. A simple coding procedure enhances a neuron’s information capacity. *Z. Naturforsch*, 36:910–912, 1981.
- [117] T.-W. Lee, M. Girolami, and T. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, in press.
- [118] T.-W. Lee, B. Koehler, and R. Orglmeister. Blind source separation of nonlinear mixing models. In *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*, pages 406–415, Florida, September 1997.
- [119] M. Lewicki and B. Olshausen. Inferring sparse, overcomplete image codes using an efficient coding framework. In M. Jordan, editor, *Advances in Neural Information Processing Systems*, volume 10, San Mateo, in press. Morgan Kaufmann.
- [120] M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural Computation*, Submitted.

- [121] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, 1993.
- [122] J. Lin, D. Grier, and J. Cowan. Source separation and density estimation by faithful equivariant som. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 536–541, Cambridge, MA, 1997. MIT Press.
- [123] R. Linsker. 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, 1986.
- [124] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
- [125] N. Logothetis and J. Pauls. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288, 1995.
- [126] D. Mackay. Maximum likelihood and covariant algorithms for independent component analysis. Unpublished manuscript obtainable at <http://wol.ra.phy.cam.ac.uk/mackay>, 1996.
- [127] D. Macleod and T. von der Twer. Optimal nonlinear codes. Technical Report 28/96, Universität Bielefeld, Zentrum für interdisziplinäre Forschung, 1996.
- [128] S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski. Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, Cambridge, MA, 1996. MIT Press.
- [129] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions E*, 74(10):3474–3483, 1991.
- [130] D. Mastronarde. Correlated firing of retinal ganglion cells. *Trends in Neuroscience*, 12(2):75–80, 1989.
- [131] M. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. Kindermann, A. Bell, and T. Sejnowski. Analysis of fmri data by decomposition into independent components. *Proc. Nat. Acad. Sci.*, in press.
- [132] M. Meister, R. Wong, D. Baylor, and C. Shatz. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–43, 1991.
- [133] K. Miller, J. Keller, and M. Stryker. Ocular dominance column development: analysis and simulation. *Science*, 245(4918):605–15, 1989.
- [134] R. Millward and A. O’Toole. Recognition memory transfer between spatial frequency analyzed faces. In H. Ellis, M. Jeeves, F. Newcombe, and A. Young, editors, *Aspects of Face Processing*, pages 34–44. Nijhoff, Dodrecht, 1986.

- [135] Y. Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(27):817–820, 1988.
- [136] P. Montague, J. Gally, and G. Edelman. Spatial signaling in the development and function of neural connections. *Cerebral Cortex*, 1:199–220, 1991.
- [137] J. Morris, C. Frith, D. Perrett, D. Rowland, A. Young, A. Calder, and R. Dolan. A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, 383(6603):812–5, 1996.
- [138] J. Movellan. Visual speech recognition with stochastic networks. In G. Tesauero, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge, MA, 1995.
- [139] D. Mumford. Pattern theory: a unifying perspective. In D. Knill and W. Richards, editors, *Perception as Bayesian inference*. Cambridge University Press, 1996.
- [140] P. N. and E. Rolls. Transform invariant recognition by association in a recurrent network. *Neural Computation*, in press.
- [141] J.-P. Nadal and N. Parga. Non-linear neurons in the low noise limit: a factorial code maximizes information transfer. *Network*, 5:565–581, 1994.
- [142] S. Nowlan. Maximum likelihood competitive learning. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 2, pages 574–582, San Mateo, CA, 1990. Morgan-Kaufmann.
- [143] K. Obermayer and G. Blasdel. Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience*, 13(10):4114–29, 1993.
- [144] K. Obermayer, G. Blasdel, and K. Schulten. Statistical-mechanical analysis of self-organization and pattern formation during development of visual maps. *Physical Review A*, 45(10):7568–89, 1992.
- [145] E. Oja. A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, 15:267–273, 1982.
- [146] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.
- [147] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
- [148] B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–340, 1996.
- [149] A. Oppenheim and J. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69:529–541, 1981.

- [150] R. O'Reilly and M. Johnson. Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, 6:357–389, 1994.
- [151] A. O'Toole, H. Abdi, K. Deffenbacher, and D. Valentin. Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, 10(3):405–411, 1993.
- [152] A. O'Toole, K. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other race effect. *Memory and Cognition*, 22(2):208–224, 1994.
- [153] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.
- [154] G. Palm. On associative memory. *Biological Cybernetics*, 36:19–31, 1980.
- [155] B. Pearlmutter and G. Hinton. G-maximization: An unsupervised learning procedure for discovering regularities. In J. Denker, editor, *Neural Networks for Computing: American Institute of Physics Conference Proceedings*, volume 151, pages 333–338, 1986.
- [156] P. Penev and J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.
- [157] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.
- [158] D. Perrett, A. Mistlin, and A. Chitty. Visual neurones responsive to faces. *Trends in Neuroscience*, 10:358–364, 1989.
- [159] M. Phillips, A. Young, C. Senior, C. Brammer, M. Andrews, A. Calder, E. Bullmore, D. Perrett, D. Rowland, S. Williams, A. Gray, and A. David. A specific neural substrate for perceiving facial expressions of disgust. *Nature*, 389:495–498, 1997.
- [160] P. Phillips, H. Wechsler, J. Juang, and P. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, in press.
- [161] G. Pike, R. Kemp, N. Towell, and K. Phillips. Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, 4(4):409–437, 1997.
- [162] L. Piotrowski and F. Campbell. A demonstration of the visual importance and flexibility of spatial-frequency, amplitude, and phase. *Perception*, 11:337–346, 1982.
- [163] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990.
- [164] D. Pollen and S. Ronner. Phase relationship between adjacent simple cells in the visual cortex. *Science*, 212:1409–1411, 1981.

- [165] J. Porrill and J. Stone. Independent components analysis for signal separation and dimension reduction. Technical Report 124, Psychology Department, Sheffield University, Sheffield, England, August 1997.
- [166] W. Pratt. *Digital Image Processing*. Wiley, New York, 1978.
- [167] P. Rhodes. 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, 1992.
- [168] W. E. Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95(1):52–77, 1984.
- [169] R. Rison and P. Stanton. Long-term potentiation and n-methyl-d-aspartate receptors: foundations of memory and neurologic disease. *Neuroscience and Biobehavioral Reviews*, 19(4):533–52, 1995.
- [170] E. Rolls. 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, 1992.
- [171] E. Rolls. Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research*, 66:177–185, 1995.
- [172] E. Rolls, G. Baylis, and M. Hasselmo. 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, 1987.
- [173] M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, 1996.
- [174] D. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9:75–112, 1985.
- [175] J. Russell and J. Fernandez-Dols. *The Psychology of Facial Expression*. Cambridge University Press, New York, 1997.
- [176] M. Rydfalk. *CANDIDE: A parametrized face*. PhD thesis, Linköping University, Department of Electrical Engineering, Oct. 1987.
- [177] T. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473, 1989.
- [178] N. Schraudolph and T. Sejnowski. Competitive anti-hebbian learning of invariants. In J. Moody, S. Hanson, and R. Lippman, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 1017–1024, San Francisco, 1992. Morgan Kaufmann.

- [179] C. Shannon and W. Weaver. *The Mathematic Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.
- [180] A. Shashua. *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, Massachusetts Institute of Technology, 1992.
- [181] R. Shepard and L. Cooper. *Mental Images and their Transformations*. MIT Press, Cambridge, MA, 1982.
- [182] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2-5 1997.
- [183] W. Singer. The formation of cooperative cell assemblies in the visual cortex. *Journal of Experimental Biology*, 153:177–197, 1990.
- [184] A. Singh. *Optic Flow Computation*. IEEE Computer Society Press, Los Alamitos, CA, 1991.
- [185] E. Steimer-Krause, R. Krause, and G. Wagner. Interaction regulations used by schizophrenic and psychosomatic patients: Studies on facial behavior in dyadic interactions. *Psychiatry*, 53:209–228, 1990.
- [186] J. Stone. Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–92, 1996.
- [187] J. Stone. Object recognition using spatiotemporal signatures. *Vision Research*, 38(7):947–951, 1998.
- [188] J. Stone and J. Porrill. Combining spatial and temporal ica to extract correlated sources. Psychology Department, Sheffield University, Sheffield, S10 2UR, England, submitted.
- [189] M. Stryker. Activity-dependent reorganization of afferents in the developing mamalian visual system. In D. Lam and C. Schatz, editors, *Development of the Visual System*, pages 267–287. MIT Press, Cambridge, MA, 1991.
- [190] M. Stryker. Temporal associations. *Nature*, 354(14):108–109, 1991.
- [191] M. Stryker and W. Harris. Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *Journal of Neuroscience*, 6:2117–2133, 1986.
- [192] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.
- [193] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

- [194] D. Tranel, A. Damasio, and H. Damasio. Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, 38(5):690–696, 1988.
- [195] M. Tsodyks and M. Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6(2):101–105, 1988.
- [196] J. Tukey. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29:614, 1958.
- [197] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [198] G. Turrigiano, K. Leslie, N. Desai, L. Rutherford, and S. Nelson. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–6, 1998.
- [199] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.
- [200] D. Valentin, H. Abdi, A. O'Toole, and G. Cottrell. Connectionist models of face processing: a survey. *Pattern Recognition*, 27(9):1209–30, 1994.
- [201] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 15(6):733–741, 1997.
- [202] H. Wallbott. Effects of distortion of spatial and temporal resolution of video stimuli on emotion attributions. *Journal of Nonverbal Behavior*, 15(6):5–20, 1992.
- [203] G. Wallis and P. Baddeley. Optimal, unsupervised learning in invariant object recognition. *Neural Computation*, 9(4):883–94, 1997.
- [204] G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology (Oxford)*, 51(2):167–194, 1997.
- [205] D. Weinshall and S. Edelman. A self-organizing multiple view representation of 3d objects. *Biological Cybernetics*, 1991.
- [206] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1994.
- [207] S. Yaxley, E. Rolls, and Z. Sienkiewicz. The responsiveness of neurons in the insular gustatory cortex of the macaque monkey is independent of hunger. *Physiology and Behavior*, 42(3):223–9, 1988.
- [208] A. Young, D. Hay, K. H. McWeeny, B. M. Flude, and et al. Matching familiar and unfamiliar faces on internal and external features. *Perception*, 14(6):737–746, 1985.

- [209] M. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 56:1327–1331, 1992.
- [210] R. Zajonc. The interaction of affect and cognition. In K. Scherer and P. Ekman, editors, *Approaches to Emotion*, pages 239–246. Lawrence Erlbaum, Hillsdale, NJ, 1984.
- [211] R. Zemel and G. Hinton. Discovering viewpoint invariant objects that characterize objects. In R. Lipmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, pages 299–305, San Francisco, 1991. Morgan Kaufmann.
- [212] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, 85(9):1423–1435, 1997.