Interactive-Time Vision: Face Recognition as a Visual Behavior

by

Matthew Alan Turk

B.S., Virginia Polytechnic Institute and State University (1982)
M.S., Carnegie Mellon University (1986)

Submitted to the Media Arts and Sciences Section,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September 1991

Author______________________________________
Media Arts and Sciences Section,
School of Architecture and Planning
August 9, 1991

Certified by______________________________________
Alex P. Pentland
Associate Professor, MIT Media Laboratory
Thesis Supervisor

Accepted by______________________________________
Stephen Benton
Chairman, Departmental Committee on Graduate Students

Interactive-Time Vision:
Face Recognition as a Visual Behavior

by

Matthew Alan Turk

Submitted to the Media Arts and Sciences Section,
School of Architecture and Planning
on August 9, 1991, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis describes a vision system which performs face recognition as a special-purpose visual task, or "visual behavior". In addition to performing experiments using stored face images digitized under a range of imaging conditions, I have implemented face recognition in a near-real-time (or "interactive-time") computer system which locates and tracks a subject's head and then recognize the person by comparing characteristics of the face to those of known individuals. The computational approach of this system is motivated by both biology and information theory, as well as by the practical requirements of interactive-time performance and accuracy. The face recognition problem is treated as an intrinsically two-dimensional recognition problem, taking advantage of the fact that faces are normally upright and thus may be described by a small set of 2-D characteristic views. Each view is represented by a set of "eigenfaces" which are the significant eigenvectors (principal components) of the set of known faces. They form a holistic representation and do not necessarily correspond to individual features such as eyes, ears, and noses. This approach provides for the ability to learn and later recognize new faces in an unsupervised manner. In addition to face recognition, I explore other visual behaviors in the domain of human-computer interaction.

Thesis Supervisor: Alex P. Pentland
Associate Professor, MIT Media Laboratory

Acknowledgments

A Ph.D. thesis represents not just a body of research accomplished over a few years of graduate school, but also the culmination of a significant period of one's life. Many people have had a significant influence on me during my time at MIT, in a variety of ways, both academic and personal. Barring an unexpected Academy Award, this may be my only opportunity to publically thank you. To all of you I express my sincere gratitude, and I hope that I can repay you in some small ways as I am able.

I would like to acknowledge in particular the continuous support of my parents and family. They have been a consistent encouragement for me throughout my many years of university education. Thanks for kindling in me an interest in scholarly things — as well as not-so-scholarly things — from the start. I'm also grateful to my parents for encouraging me to enjoy music, sports, literature, and other parts of life as well as science, engineering, and technology.

Ron, Kevin and Stacy have been invaluable in their long-term friendships, their encouragement, and their ability to make me both laugh and think. Thanks for helping to keep my spirits up over the years.

My four years at MIT have been priceless. The Media Lab is a gem — the people, the facilities, the freedom I've had, and the general MIT environment are unique. Hats off to Nicholas Negroponte, Jerome Wiesner, and the others who brought the Lab to life. Sandy Pentland and Ted Adelson have done a superb job in getting together a top-notch vision research group in just a few years, and I'm quite impressed that great scientists can also be great managers and organizers. You both have my upmost respect. To all of the Vision and Modeling Group folks — I've really enjoyed it! Thanks to Laureen and Bea for loads of help and for laughing at my silly jokes. To Mike, Bill, John M., and Irfan, my officemates at various times, for putting up with an often messy office and for just hanging out at times. Thanks to Eero, Dave, Mike, and others who have contributed to OBVIUS, for building a tool that I could complain about for years — while at the same time depending on it and actually liking it! Trevor rescued my kamikaze files and has graciously answered my many cries for system help. Stan, Bradley, Marty, Roz, Kenji, Stephen, Gianni, Stanzi, Thad, Pawan, and others have in various ways made this a great place to be, both intellectually and personally. Thanks also to the other Media Lab grad students and friends and colleagues at the AI Lab and the Brain and Cognitive Science Department for sharing your knowledge and insight.

For all of you who let me "borrow" your face, or uncomplainingly complied with my requests to digitize you “one more time”, I owe you one!

I want to thank the Arbitron Company for supporting my work, and particularly Tony Gochal for believing in it and for being a pleasure to work with.

Thanks to my committee, Sandy Pentland, Ted Adelson, and Nicholas Negroponte, for fitting me into such busy schedules and coping with my last-minute nature, and for detailed comments that have improved this document immensely.

Special thanks go to Sandy Pentland, who has been not just an advisor but a colleague and friend as well. I could not have asked for a better thesis advisor — he has provided a good balance of freedom and interest, has been a constant source of ideas and suggestions, and has kept me free from funding concerns. And many thanks to Sandy and Ted for encouraging me to come to the Lab in the first place.

Many friends have contributed to keeping my sanity while in graduate school, including the Club Cornelius guys (and Ladies Auxiliary), the folks at *pika*, the L'Abri regulars, Park Street folks, the Shadowmasks, and my basketball cohorts. (In my heart I'll always be “Slow But Short”!)

The best thing to happen to me during my time at MIT has nothing to do with a Ph.D. My gratitude to my fiance Kelly will take a lifetime to express, but fortunately we'll have just that long together. I love you, and look forward to sharing life with you.

This research was performed at the MIT Media Laboratory and was supported in part by Control Data Corporation under contract number M87-M114, and in part by the USAF Rome Air Development Center under contract number F30602-89-C-0022.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Most working robots also cannot see. If a part is missing from an assembly line, a robot will act as if it were there. In the future, robots will be able to "see" as well as sense pressure. They will do more jobs and even better ones than they do today.*
>
> *C-3PO's Book About Robots*

1.1 Face recognition by man and machine

The human ability to recognize faces is remarkable. Faces are complex visual stimuli, not easily described by simple shapes or patterns; yet people have the ability to recognize familiar faces at a glance after years of separation. The subject of visual processing of human faces has received attention from philosophers and scientists for centuries. Aristotle devoted six chapters of the *Historia Animalium* to the study of facial appearance. Physiognomy, the practice or art of inferring intellectual or character qualities of a person from outward appearance, particularly the face, has had periods of fashion in various societies [62]. Darwin considered facial expression and its identification to be a significant advantage for the survival of species [28]. Developmental studies have focused on strategies of recognition or identification and the differences between infant and adult subjects. Neurological disorders of face perception have been isolated and studied, providing insight into normal as well as abnormal face processing. In recent years, computers have been introduced into various aspects of the subject, with systems which attempt to model, display, and recognize human faces.

There is something about the perception of faces that is very fundamental to the human experience. Early in life, we learn to associate faces — particularly the mother's face — with pleasure, fulfillment, and security. As we get older, the subtleties of facial expression — a glance between friends or lovers, a chilling stern look from one in authority — enhance our explicit communication in myriad ways. The face is our primary focus of attention in social intercourse; this is observable in interaction among animals as well as between humans and animals. The face, more than any other part of the body, communicates identity, emotion, race, and age, and also is quite useful for judging gender, size, and perhaps even character.

A desire to understand and to replicate (or at least approximate) the human ability to recognize and "read" faces has naturally lead to the advent of computational approaches to face processing. Computer systems that can recognize and identify faces may be useful in a number of applications. For example, the ability to model a face and distinguish it from a large number of stored face models is essential for the automation of criminal identification. The non-intrusive nature of face identification is well suited for security systems. The detection of faces in photograph negatives or originals will be quite useful in color film development, since the effect of many enhancement or noise reduction techniques depends on the picture content. Automated color enhancement is desirable for most parts of the scene, but may have an undesirable effect on flesh tones. (It is fine for the yellowish grass to appear greener, but not so fine for Uncle Harry to look like a Martian!) In the areas of image compression for transmission of movies and television, and in general any "semantic understanding" of video signals, the presence of people in the scene is important. For example, in partitioning the spatial-temporal bandwidth for an advanced HDTV transmission, more bandwidth should be given to people than to cars, since the audience is much more likely to care about the image quality and detail of the human actors than of inanimate objects.

In the area of human-computer interaction, an ultimate goal is for machines to understand, communicate with, and react to humans in natural ways. A machine that understands gestures, perceives direction of gaze, uses lip reading to disambiguate and facilitate speech recognition, and visually identifies individuals and their moods and emotions, is the stuff of science fiction — e.g. the computer Hal in *2001: A Space Odyssey* — yet all these problems are currently being addressed by a number of researchers. Although there are many other avenues to person identification — gait, clothing, hair, voice, and height are all useful indicators of identity — none are as compelling as face recognition.

In computer graphics and related fields, the accurate graphic display and animation of faces has many applications as well. There are research groups working towards animated actors in film and video productions. A computer system processing face information has recently been publicized through television news and popular magazine reports [68]. This system simulates aging by applying appropriate transformations to facial images, and has been directly responsible for the location of missing children years after their disappearance. In the medical field, surgeons and computer scientists are working to develop interactive graphic systems for modeling and prediction in craniofacial surgery. Coding faces for low-bandwidth teleconferencing is also an active area of research.

As these observations suggest, the visual processing of faces is an interesting and potentially useful and enlightening direction of research. The goals are quite ambitious, in that these are high-level visual tasks, while basic areas such as stereo and motion perception are still not completely understood. However, the tasks involved in face processing are reasonably constrained; some may even have a degree of "hardwiring" in biological systems. Faces present themselves quite consistently in expected positions and orientations; their configuration (the arrangement of the components) seldom changes; they are rather symmetrical. On the other hand, human face recognition and identification is very robust in the face of external changes (e.g. hair styles, tan, facial hair, eyeglasses), so a recognition scheme cannot be overly constrained.

1.2 Paradigms of visual object recognition

Models of recognition have been debated in neuroscience for decades. The oversimplified view of object recognition as hierarchical feature extraction [74] involves cells at different stages of processing signaling different visual features. At the earliest level, cells detect basic image features such as edges at particular orientation, position, and contrast. At higher stages neurons are tuned to combinations of these basic features, corresponding to more complex features or shapes. At the apex of the hierarchy are so-called "gnostic units", whose activation signal the perception of familiar objects over different viewing conditions. The infamous "grandmother cell", which fires whenever one's grandmother is in the visual scene, is an example of a gnostic unit. With the highest level of gnostic cells replaced by the concept of population encoding of neurons, this is a popular neural model of visual processing, at least in neuroscience textbooks.

The reigning paradigm for object recognition in computational vision involves matching stored models to representations built from the image data, beginning with the detection of features and moving on to some description of surfaces or parts. This modular approach to vision, often referred to as the *Marr paradigm* [64], involves some variation of the following framework (see also Figure 1-1(a)):

1. The primal sketch or intrinsic images: From the input image intensities, make important information (e.g. intensity changes and their 2-D geometry) explicit.
2. $2\frac{1}{2}$-D sketch: Calculate the orientation, depth, and discontinuities of visible surfaces in a viewer-centered frame.
3. 3-D representation: Describe the object shape, using volumetric and surface primitives in an object-centered frame.

Each step of the Marr paradigm involves a transformation to a "higher-level" representation, one that reduces the imaging-specific dependencies such as illumination and viewpoint. Recognition is possible once the image data and the object models are in a common form. At the highest level, the 3-D shape representation is compared to models of known 3-D objects.

In the past decade, much effort has been devoted to discovering useful constraints and exploring ways to make this paradigm computationally feasible. The use of groupings such as those proposed by the Gestalt psychologists [63], multiple canonical representations [71], and alignment techniques [102] have encouraged a degree of interaction between previously independent levels of the recognition process.

While this approach is appropriate for general object recognition, there appears to be both biological and computational motivation to develop special-purpose recognition capabilities. Bruce and Young [15], for example, argue that Marr's representation scheme is not suitable to cope with the fine discriminations needed for face recognition, where all the "objects" have similar overall shapes.

In recent years a new paradigm has gained popularity in the field of computer vision which emphasizes fast, goal-oriented vision (e.g. [96, 3, 4]). These ideas, termed

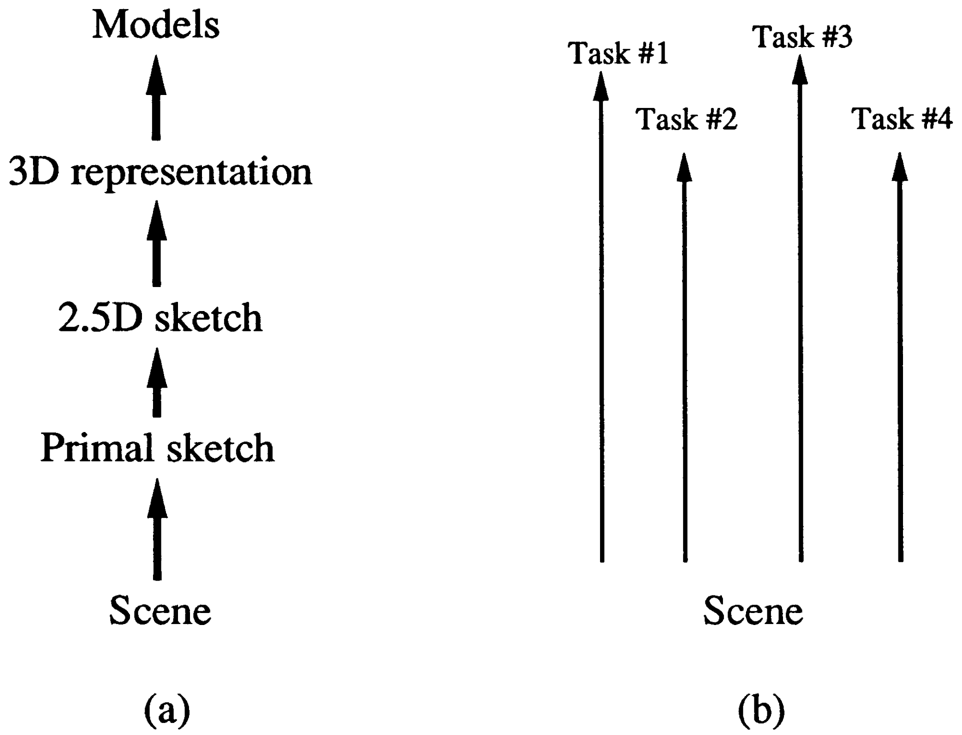


Figure 1-1: (a) The Marr vision paradigm: horizontal layers (b) The active vision, or visual behaviors, paradigm: vertical layers

active vision or *animate vision*, are in most aspects orthogonal to the more general Marr paradigm, emphasizing relatively simple processing and recognition strategies in limited domains. While Marr viewed the main job of vision to be to derive a representation of shape, the function of an active vision system is application-dependent.

1.3 Visual behaviors

In his article, "Intelligence without representation," Rod Brooks [13] argues for developing intelligent systems incrementally, relying on interaction with the real world rather than on representations. He describes limitations of the traditional notion among artificial intelligence researchers of functional decomposition (e.g. the Marr paradigm), and proposes a fundamental slicing up of intelligent systems into *activity* producing subsystems, each of which is best suited for a certain "ecological niche". The advantage of approaching intelligence through such layers of activity, or skills, or behaviors, is that "it gives in incremental path from very simple systems to complex autonomous intelligent systems." These multiple parallel activities do not depend on central representations; instead, the collection of competing behaviors can produce a coherent pattern of behavior similar to Minsky's [66] theories of human behavior.

The distinctions between Brooks' views and the traditional AI views of intelligence are parallel to the paradigms of object recognition discussed above. They are not *either/or* alternatives, but instead different ways of solving different problems. Biological vision gives evidence for both general and special purpose components — a likely proposition is that in evolutionary terms, early visual systems are primarily special-purpose and later visual systems added capabilities such as general object recognition. Figure 1-1(b) shows a schematic representation of the active vision, or visual behaviors paradigm.

I claim that the ability to quickly recognize faces is primarily a special-purpose system biologically, and that it makes sense to approach it as a special-purpose system (a visual behavior or active vision component) computationally. Human faces all share the same basic configuration and 3-D structure — the subtle differences are what distinguish one from another. While it is certainly possible to inspect and compare face shapes and complex structure, the physiological and psychological studies of face perception (see Chapter 2) seem to point to a fast, special-purpose face recognition system somewhat independent of general recognition.

Although there is no compelling reason to model intelligent computer-based systems directly on biological systems, the existence proof of the biological system is at the least suggestive. As is often pointed out, modern airplanes do not flap their wings as birds, yet the study of birds and flight led to discovering the *principles* of aerodynamics which enables planes to fly. Intelligent computer systems do not need to duplicate biological strategies, but to learn from their lessons. There is sufficient motivation to devote research energy into the "visual behaviors" approach to object recognition, and face recognition in particular.

1.4 Face recognition and interactive-time vision

Figure 1-2 depicts the intersection of three aspects of vision research which are applicable to the pursuit of computer-based face recognition. In human-computer interface (HCI), we want to model human-to-computer interaction after human-to-human interaction, to the degree that it is appropriate. Menu-based applications and graphical user interfaces (GUIs) just begin to scratch the surface of "user-friendly" machines. Before people will really accept the idea of intelligent machines being a vital part of society, and before these machines are really accessible to the general public, the man-machine interaction must be considered natural. This will involve not only currently important technologies such as speech recognition and synthesis, but more subtle abilities such as recognizing gestures, identity, and facial expressions from visual input.

"Interactive-time vision" is a term meant to include not only real-time systems (i.e. systems which give an answer in a matter of milliseconds or less, fast enough for the inner control loop of some process), but also systems which respond quickly enough for the task at hand to feel interactive to the user or observer. As with the "real-time" label, the boundaries of interactive-time are always changing along with improvements in hardware and software. However it is not just a label of convenience — people demand interactive-time performance in order to consider a machine an active, intelligent participant rather than an advanced but impersonal calculator. In addition, interactive-time systems can be tested *in the real world*, on real time-varying data, in a way that non-interactive systems cannot. Rather than limiting the environment to overly simple scenarios (e.g. the "blocks world" in vision), or to stored unreal data (e.g. synthetic scenes plus gaussian noise), significant advances should be made by developing intelligent systems which can interact with the real,

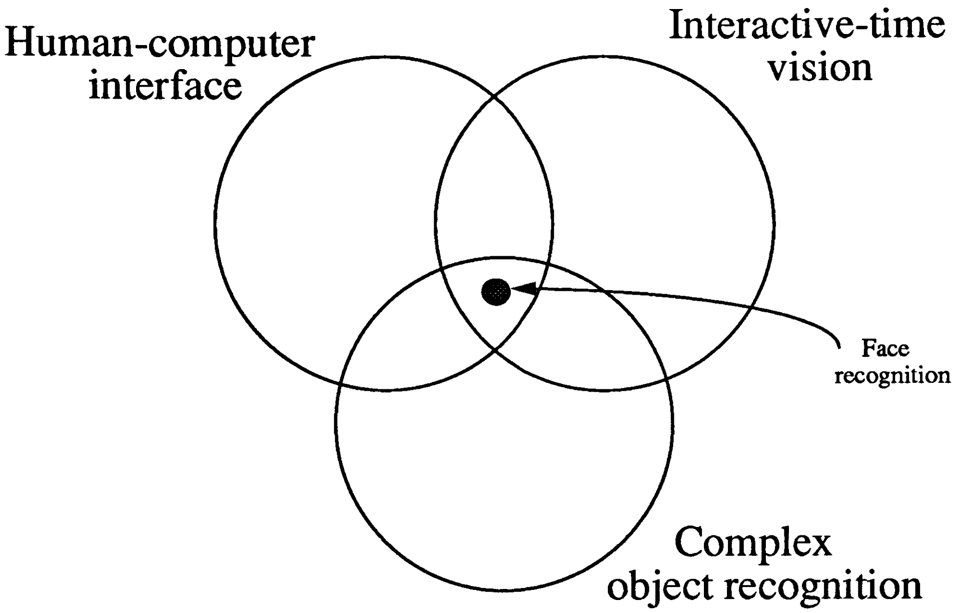


Figure 1-2: Face recognition lies in the intersection of three fruitful areas of research: human-computer interface, interactive-time vision, and complex object recognition.

dynamic world.

Similar to Brooks' requirements for his "Creatures", or autonomous agents[13], requirements for an interactive-vision system include the following:

1. The system must act appropriately and timely in a dynamic environment.

2. The system should be robust, in that small changes in the environment should lead to at worst a gradual decline in the system performance rather than a total collapse.

3. The system should be able to maintain multiple goals — e.g. a face recognition system (or behavior) may be used to either locate faces, identify them, or both.

4. The system should have some useful (behavior-enabling) skill on its own. (So for example a stereo module which produces a range map as its output does not qualify).

Faces are complex objects, not easily described by simple features, surface models, or volumetric models. Trees, water, and clouds are examples of other complex objects which can be modeled relatively well for rendering in computer graphics, but cannot be reliably recognized by any current computer vision system. The strategy of building increasingly complex systems to recognize complex objects is not the only option. It is possible at times — as in the case of faces — to develop relatively simple strategies and rely on rather simple representations to support recognition in some limited domain of circumstances. It may be more efficient, for example, to build "oak tree detectors" and "maple tree detectors" rather than general "tree detectors".

My approach to face recognition lies in the intersection of these three components of interest — human-computer interface, interactive-time vision, and complex object recognition. Going along with the "visual behaviors" idea, it seemed fruitful to approach the practical problem of recognizing human faces in this framework. The research of this thesis shows this to be the case.

1.5 Overview of the thesis

This research has been focused towards developing a sort of early, preattentive pattern recognition capability that does not depend upon having three-dimensional information or detailed geometry. The goal is to develop a computational model of face recognition which is fast, reasonably simple, and accurate in constrained environments such as an office or a household. In addition the approach is biologically implementable and is generally in concert with preliminary findings in the physiology and psychology of face recognition.

The scheme is based on an information theory approach that decomposes face images into a small set of characteristic feature images, called "eigenfaces", which are the principal components of the initial training set of face images. Recognition is performed by projecting a new image into the subspace spanned by the eigenfaces ("face space") and then classifying the face by comparing its location in face space with the locations of known individuals.

Automatically learning and later recognizing new faces is practical within this framework. Recognition under widely varying conditions is achieved by training on a limited number of characteristic views (e.g., a frontal view, a 45° view, and a profile view). The approach has advantages over other face recognition schemes in its speed and simplicity, learning capacity, and insensitivity to small or gradual changes in the face image.

Chapter 2 surveys the relevant literature on biological and computational face recognition. Some insight from the physiological studies of face-selective cells, as well as studies on disorders of face recognition abilities and the strategies people use in recognizing faces, was motivational to the computational approach presented in this thesis.

Chapter 3 introduces a method of computational face recognition using "eigenfaces", while Chapters 4 and 5 describe experiments exploring the performance of the approach on stored face images and with a near-real-time system. The biological implications of the system and some relation to neural networks are discussed in Chapter 6.

The general class of interactive-time systems performing useful human-computer interface tasks is explored further in Chapter 7, where a simple system to detect eye blinks — and therefore alert observers — is described, as well as investigations into detectors for the direction of gaze and expressions. Chapter 8 summarizes the main ideas of the thesis and the state of the current implementation, and discusses future research directions and work in progress at other labs building on these ideas. Appendix A relates part of this work to the techniques of correlation and matched filtering.

Before continuing, a brief note about terminology. The terms "recognition" and "identification" are often used interchangeably in both common conversation and in the scientific literature. In this thesis they are also quite often both used to mean the general process of perceiving, and perhaps establishing the identity of, a known object. When it is necessary to distinguish between different aspects of this process — which will be evident in the context — *recognizing* a face will refer to the perception of a face as a face, while *identifying* a face will refer to correctly naming the perceived face as one out of a known group. These distinctions are unfortunately often ignored in both the human vision and computer vision literature, although in principle there may be quite different mechanisms underlying the two.

Chapter 2

Background

> *CHURCH-TURING THESIS, THEODORE ROSZAK VERSION:*
> *Computers are ridiculous. So is science in general.*
>
> Douglas Hofstadter, *Gödel, Escher, Bach*

2.1 Introduction

In the past two decades there has been a growing interest in face recognition and identification in physiology, neurology, psychology, and computer vision. Motivated by such diverse interests as commercial security systems and people meters, model-based coding for telecommunications, understanding the development of human visual capabilities from infant to adult, and understanding visual dysfunction in brain-damaged patients, face recognition has become a popular topic. An understanding of the processes involved in face recognition may reveal important clues as to the neural structures underlying recognition and important hints to the construction of computational recognition systems. This chapter reviews the relevant literature in both biological and computational vision.

2.2 Prosopagnosia

Visual agnosia is a neurological impairment in the higher visual processes which leads to a defect in object recognition [34]. Agnosic patients can often "see" well, in that there is little apparent deficit in spatial vision or perception of form. The dysfunction is specific to some class of objects or shapes, such as perceiving letters or any object from an unusual viewpoint. Etcoff *et al.* [38] report a patient's description of his agnosia to be like "attempting to read illegible handwriting: you know that it is handwriting, you know where the words are and letters stop and start, but you have no clue as to what they signify."

Lissauer's seminal paper on visual agnosia in 1890 (see [90] for an abridged English version with commentary) presented the first thorough clinical description of an agnosic patient, and distinguished between two aspects or forms of agnosia: *apperceptive* and *associative*. Apperception is the process of constructing a perceptual representation from the visual input, while association is the process of mapping a perceptual representation onto stored knowledge of the object's functions and associations [80]. So apperceptive agnosia involves a problem in constructing the perceptual representation, while in associative agnosia there is difficulty associating the representation with any memory of the specific object.

Prosopagnosia, from the Greek *prosopon* (face) and *agnosia* (not knowing), refers to the inability to recognize familiar faces by visual inspection[1] [37, 83, 27]. Prosopagnosics can typically identify the separate features of a face, such as the eyes or mouth, but have no idea to whom they belong. They may recognize the sex, age, pleasantness, or expression of a face, without an awareness of the identity:

> I was sitting at the table with my father, my brother and his wife. Lunch had been served. Suddenly... something funny happened: I found myself unable to recognize anyone around me. They looked unfamiliar. I was aware that they were two men and a woman; I could see the different parts of their faces but I could not associate those faces with known persons.... Faces had normal features but I could not identify them. [Agnetti et al., p. 51, quoted in [29]]

Studies of covert recognition (e.g. [98, 30]) show that some prosopagnosics actually carry out some steps of the recognition process despite their lack of awareness, leading to the suspicion that prosopagnosia is an associative agnosia. However, others show no signs of covert recognition. Tranel and Damasio [98] suggest a four-part model of facial learning and recognition:

[1] Using the precise terminology described at the end of Chapter 1, prosopagnosics can *recognize* faces but not *identify* them. But I defer here to the terminology of the sources.

1. Perception

2. Templates — records of past visual perceptions of a face can be aroused by current perception.

3. Activation — multimodal memories corresponding to the face are evoked.

4. Conscious readout — the experience of familiarity.

They suggest that impairment of the activation step may explain prosopagnosia. As we will see in Chapter 3, these parts loosely correspond to my computational approach to face recognition. Prosopagnosic patients, although very few in number, have proved to be a valuable resource in probing the function of face recognition.

2.3 Face-selective cells

There is evidence that damage to the a particular area of the right hemisphere has a predominant role in producing face recognition difficulties. The question arises, is face recognition a special, localized, subsystem of vision?

One way to approach this question, and additionally to learn about the neural mechanisms involved in face recognition and object recognition in general, is by recording the activity of brain cells while performing visual tasks including observing and recognizing faces. Through single cell recording, a number of physiologists have found what seem to be "face" neurons in monkeys, responding selectively to the presence of a face in the visual field. Perrett *et al.* [72, 74, 75] have found cells in area STS of the rhesus monkey which were selectively responsive to faces in the visual field. Many of these cells were insensitive to transformations such as rotation. Different cells responded to different features or subsets of features, while most responded to partially obscured faces. Some cells responded to line drawings of faces. About 10% of the cells were sensitive to identity. Other researchers (e.g. [14, 31, 82]) have found cells with similar properties in monkey inferior temporal cortex, concluding that there may be specialized mechanisms for the analysis of faces in IT cortex. Kendrick *et al.* [52] have even found face-selective cells in sheep.

Table 2.1 lists various properties of these face cells reported by various laboratories. One should be cautious about drawing broad conclusions about face recognition from these findings. They may seem to suggest a uniqueness of face recognition, a rather localized and "hard-wired" system of grandmother-like cells. A careful look at the data, however, suggests some sort of population coding of information, and a not very straightforward one at that.

In a review article, Desimone [32] suggests that "face cells could turn out to be a model system for studying the neural mechanisms of complex object recognition, rather than an exception." Although properties of the "eigenfaces" or the recognition strategy described in Chapter 3 are analogous to many of properties of the faces cells studied to date, there is no evident one-to-one correspondence. As the face cells become better understood, they should motivate other approaches to complex object recognition as well.

2.4 Mechanisms of face recognition

Psychologists have used both normal and prosopagnosic subjects to investigate models of face processing, recognition, and identification. In addition to the theoretical and clinical pursuits of neuroscience, the validity and limitations of eye-witness testimony in criminal proceedings has spurred much face recognition research in cognitive psychology.

Yin [106] presented pictures of faces in various orientations and tested subsequent recall, finding that the recall performance for inverted faces was degraded more than that of other configuration-specific stimuli such as landscapes or animals. He argued for a special face-processing mechanism to account for this effect. Others have furthered these techniques to experiment with face images which have been modified in myriad ways.

Developmental studies (e.g. [61]) have observed the development of face recognition from infant to adult. Carey and Diamond [20] found that the effect of inversion on face recognition described by Yin increases over the first decade of life, suggesting that young children represent faces in terms of salient isolated features ("piecemeal representation"), rather than in terms of configurational properties used by older children and adults ("configurational representation"). In recent years there seems to be a growing consensus that both configurational properties and feature properties are important for face recognition [14].

Carey and Diamond [33] claim that face recognition is not a special, unique sys-

Property	**Results**
color	most cells not sensitive to color [72]
orientation	not very sensitive to orientation (rotation in the viewing plane) [31]
line drawings	most cells not significantly responsive to line drawings of faces [72]
position	not dependent on position of the face [31]
size	most not dependent on size of face [31, 81]
contrast	relatively invariant to magnitude and sign of contrast [81]
identity	about 77% respond differently to different faces [6]
identity	seems to be encoded not in individual neurons but in an ensemble [6]
identity	most responded independent of identity [72]
identity	about 10% are sensitive to identity [75]
expression	about 10% are sensitive to expression [75]
identity/expression	some are sensitive to expression but not identity [74]
identity/expression	expression and identity seem to be encoded by separate populations of cells in separate anatomical locations [32]
face view	some respond best to front view, some to profile [31]
face view	most are view-dependent [72]
face view	view-dependent cells have been identified across the entire 360° range [47]
occlusion	most respond despite occluding some parts of the face [72]
features	most respond more to the whole face than to any one part [72, 31]
features	many are sensitive to the presence of a particular facial feature only [75]
features	scrambling the configuration of features reduces or eliminates the response [31, 75]
features	cells detect the combination of distances between different parts of the face [105]
eye contact	many respond best when the observed face is making eye contact [74]

Table 2.1: Properties of face-selective cells from single-cell recordings in monkey cortex (areas STS and IT).

tem, and that the inversion effect may be due to a gain in the ability to exploit distinguishing "second-order relational features". For faces and many other complex objects the *first-order relational features* — the spatial relationships between similar parts — are constrained. Such objects must be differentiated by distinctive relationships among the elements of the common configuration. Such *second-order relational features* may be vital in many complex object recognition tasks. What is important here is that the strategies used for face recognition should be applicable to many other recognition tasks.

A number of experiments have explored feature saliency, attempting to discern the relative importance of different features or areas of the face. Although the early of these generally agreed to the importance of face outline, hair, and eyes — and the relative unimportance of the nose and mouth — there is evidence that these results may be biased by the artifacts of the techniques and face presentations used [14].

Along with stored face images, a number of researchers [11, 59] have used face stimuli constructed from Identikit or Photofit[2] to explore strategies of face recognition. Use of these kits may actually bias the experiments, however, since there is an underlying assumption that a face can be properly decomposed into its constituent features: eyes, ears, nose, mouth, etc.

One lesson from the study of human face recognition is that approaches which treat faces as a collection of independent parts are unlikely to be relevant to the perception of real faces, where the parts themselves are difficult or impossible to delimit [14]. Consequently artists' sketches are better than face construction kits in reproducing the likeness of a target face. Faces grow and develop in a way such that features are mutually constraining. In fact these growth patterns can be expressed mathematically and used to predict the effects of aging [76]. Such techniques have already been used successfully in the location of missing children years after their disappearance [68].

Other studies have shown that expression and identity seem to be relatively independent tasks [48, 107], which is also supported by some neurological studies of prospoagnosics.

[2] Identikit and Photofit are face construction kits, used mostly by police departments, which superimpose layers of facial features to produce a large number of different faces. Newer systems, such as the Minolta Montage Synthesizer, use optics or image processing to blend together a composite face image.

Caricatures are simplified yet exaggerated representations of faces — a facial coding which "seeks to be more like the face than the face itself" [12]. Studies of caricature can provide insight into the mental representation and recognition of faces. Rhodes, Brennan, and Carey [79] tested recognition ability on faces shown as photographs, line drawings, and caricatures of varying extent. Their results are consistent with a holistic theory of representation in which distinctive aspects of a face are represented by comparison with a norm, referred to as the *distinctiveness hypothesis.* The representation based on "eigenfaces" described in Chapter 3 is based on a similar notion, since only the deviation from the average face is encoded.

2.5 Computational approaches to face recognition

Much of the work in computer recognition of faces, for the last twenty-five years, has focused on detecting individual features such as the eyes, nose, mouth, and head outline, and defining a face model by the position, size, and relationships among these features. In the past decade new approaches have emerged, most notably those based on neural networks, correlation-based techniques, and shape matching from range data.

2.5.1 Feature-based approaches

Bledsoe [9, 10] was the first to report semi-automated face recognition, using a hybrid human-computer system which classified faces on the basis of fiducial marks entered on photographs by hand. Parameters for the classification were normalized distances and ratios among points such as eye corners, mouth corners, nose tip, and chin point. At Bell Labs, Harmon, Goldstein and their colleagues [39, 44] developed an interactive system for face recognition based on a vector of up to 21 features, which were largely subjective evaluations (e.g. shade of hair, length of ears, lip thickness) made by human subjects. The system recognized known faces from this feature vector using standard pattern classification techniques. Each of these subjective features however would be quite difficult to automate.

Sakai *et al.* [85] described a system which locates features in a Laplacian-filtered image by template-matching. This was used to find faces in images, but not to recognize them. A more sophisticated approach by Fischler and Elschlager [35] attempted to locate image features automatically. They described a linear embedding algorithm which used local feature template matching and a global measure to perform image matching. The technique was applied to faces, but not to recognition.

The first automated system to recognize people was developed by Kelly [51]. He developed heuristic, goal-directed methods to measure distances in standardized images of the body and head, based on edge information.

Kanade's face identification system [50] was the first automated system to use a top-down control strategy directed by a generic model of expected feature characteristics of the face. His system calculated a set of facial parameters from a single face image, comprised of normalized distances, areas, and angles between fiducial points. He used a pattern classification technique to match the face to one of a known set, a purely statistical approach depending primarily on local histogram analysis and absolute gray-scale values. Figure 2-1 shows the fiducial points and the definition of the parameter vector.

In a similar spirit, Harmon *et al.* [45, 46] recognized face profile silhouettes by automatically choosing fiducial points to construct a 17-dimensional feature vector for recognition. Others have also approached automated face recognition by characterizing a face by a set of geometric parameters and performing pattern recognition based on the parameters (e.g. [19, 25, 104]).

The local/global template matching approach by Fischler and Elschlager has been extended by the recent work of Yuille *et al.* [108, 109]. Their strategy is based on "deformable templates", which are parameterized models of features and sets of features with given spatial relations. Figure 2-1(b) shows a deformable template for an eye. The parameter values of these models are set to initial defaults, corresponding to a generic face or perhaps the expected face, and are dynamically updated by interactions with the image through a gradient descent method. Shackleton and Welsh [89] use this method for finding facial features.

2.5.2 Connectionist approaches

Connectionist, or neural network, approaches to face identification seek to capture the holistic or gestalt-like nature of the task, using the *physical systems model* of pro-

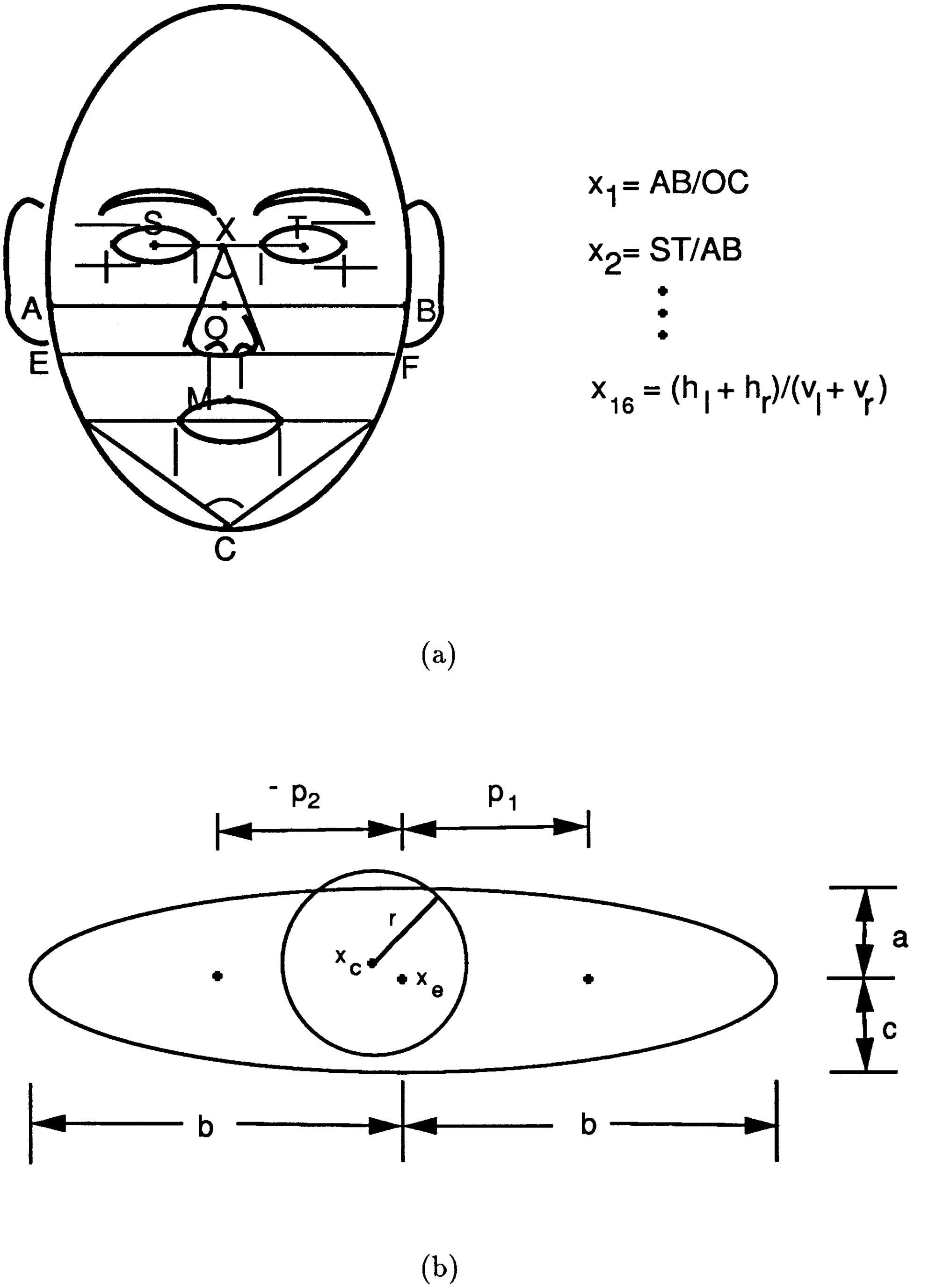


Figure 2-1: (a) Kanade's fiducial points and the corresponding face pattern vector. (From Kanade [50], reprinted with permission). (b) A deformable template of an eye. (From Yuille [109], reprinted with permission).

cessing and memory rather than the standard *information processing model* common to much vision research. Kohonen *et al.* [55, 56] demonstrated a linear associative network with a simple learning algorithm which can classify input patterns and recall a pattern from an incomplete or noisy version input to the network. Human faces were used to demonstrate the associative recall. These ideas are further investigated by O'Toole and Adbi [1, 69].

A number of researchers (e.g. [65, 49]) have used faces or face features as input and training patterns to networks with a hidden layer, trained using backpropagation, but on small data sets. Fleming and Cottrell [36] extend these ideas using nonlinear units, training the system by back propagation. The system accurately evaluated "faceness", identity, and, to a lesser degree, gender, and reported a degree of robustness to partial input and brightness variations. Cottrell and Metcalfe [24] build on this work, reporting identity, gender, and facial expression evaluations by the network.

The WISARD system [94] is a general-purpose binary pattern recognition device based on neural net principles. It has been applied with some success to face images, recognizing both identity and expression.

2.5.3 Range-based approaches

Range data has the advantage of being free from many of the imaging artifacts of intensity images. Surface curvature, which is invariant with respect to viewing angle, may be quite a useful property in shape matching and object recognition. Lapresté *et al.* [58] present an analysis of curvature properties of range images of faces, and propose a pattern vector comprised of distances between characteristic points. Sclaroff and Pentland [87] report preliminary recognition results based on range data of heads.

Lee and Milios [60] explored matching range images of faces represented as extended gaussian images. They claim that meaningful features correspond to convex regions and are therefore easier to identify than in intensity images. Gordon [42] represents face features based on principal curvatures, calculating minimum and maximum curvature maps which are used for segmentation and feature detection.

The major drawback of these approaches is the dependency on accurate, dense range data, which is currently not available using passive imaging systems and very cumbersome and expensive using active systems. In addition, it is not clear that range information alone is sufficient for reliable recognition [15].

2.5.4 Other approaches

A number of computational approaches to face recognition do not fit comfortably under any of the above labels. Baron [5] described a correlation-based approach which used template-matching to locate the eyes and subsequently to recognize and verify the face. The face recognition work by Burt *et al.* uses a "smart sensing" approach [16, 17, 18] based on multiresolution template matching. This coarse-to-fine search strategy uses a representation of the face called a pattern tree, where distinctive patterns are represented represented in more detail than the complete head. It is implemented on a special-purpose computer built to calculate multiresolution pyramid images quickly, and has been demonstrated identifying people in near real-time. The face models are built by hand from single face images.

Sakaguchi sl et al. [84] propose face identification using isodensity (or isointensity) images, in which people are identified by comparing the shape of isodensity lines. Isodensity lines are related to local orientation, as the orientation at any image position is orthogonal to the isodensity line passing through that point. Bichsel [8] has developed a face recognition system based on matching feature templates of local orientation. Local orientation should be more reliable than intensity information because of its relative invariance to contrast.

Some recent systems attempt to locate, but not identify, faces in images based on simple models of skin color [86], face outlines [43], or feature and shape "experts" [97].

2.6 Observations

All of the face recognition systems or approaches mentioned in this chapter (and others not mentioned) have one thing in common: they do not perform general, unconstrained, interactive-time recognition of faces. All are limited in their ability to perform under varying condition of lighting, scale, and viewpoint, and with facial changes due to expression, aging, and facial hair. Although recognition rates of up to 99.2% are reported, the numbers are only marginally meaningful without understanding the relative limitations of the techniques.

This is not meant to be a disparaging remark, but rather an occasion for thinking about appropriate directions of investigation in the field. It seems that the short history of research in computational face recognition is one of poking around in the "space of possible approaches", and occasionally finding promising areas to pursue. The surging of interest in this research area in recent years raises the question, discussed in Chapter 1, of whether we should direct efforts toward developing modules of increasing generality (a functional decomposition) or developing systems which work in limited domains (a behavior-based approach). My approach presented in the next chapter is both an example of continued poking — *What new functions need to be explored?* — and an attempt to put together some useful ideas gleaned from disparate approaches into a working system, within the behavior-based (or "visual behavior") framework.

Although studies of face recognition in physiology, neurology, and psychology provide little practical guidance for computer vision systems at this point, they nonetheless provide insight into the problem. While the approach taken in this thesis is not an attempt to model human strategies or biological solutions to face recognition and identification, many of its components were motivated by the human vision literature. These will be discussed further in Chapter 6.

Chapter 3

Recognition Using Eigenfaces

Have you ever watched your friend asleep — to discover what he looked like? Yet your friend's face is something else beside. It is your own face, in a rough and imperfect mirror.

Friedrich Nietzsche, *Thus Spoke Zarathustra*

3.1 Introduction

Consonant with the goals of interactive-time vision discussed in Chapter 1, the objectives of this approach to face recognition include:

- Speed — the online processing must be fast, i.e. reasonably simple mechanisms.
- Accuracy — recognition performance must be high, as compared with other approaches, and possibly tunable for the intended application.
- The system should be robust with respect to noise, variations in imaging conditions, occlusions.
- Learning — There should be some capacity to learn new faces in an unsupervised manner.
- The tasks of finding faces and identifying them should be separately achievable goals.

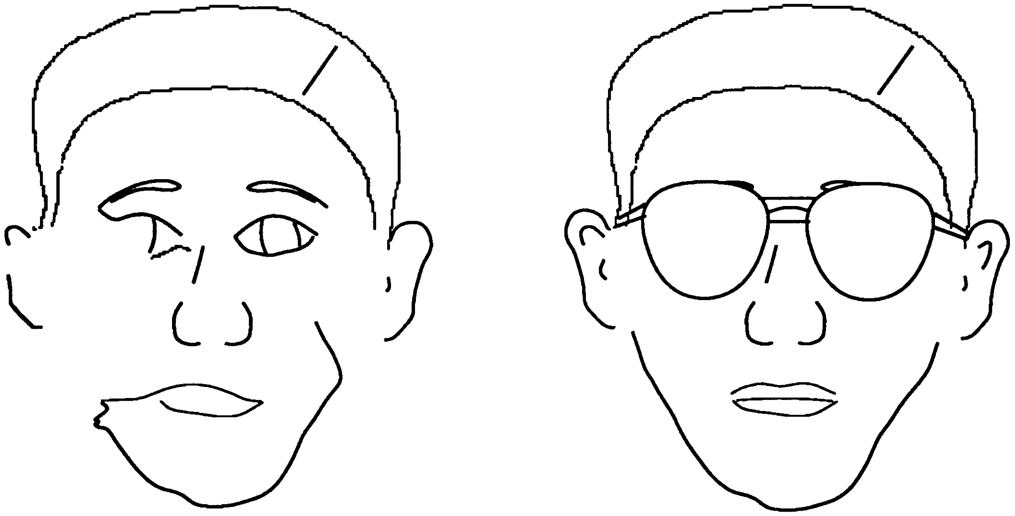


Figure 3-1: A feature-based recognition approach may be difficulty dealing with situations such as these: noisy data (resulting in missing or bad edge information) or sunglasses occluding the eyes.

Much of the previous work on automated face recognition has ignored the issue of just what aspects of the face stimulus are important for identification, by either treating the face as a uniform pattern or assuming that the positions of features are an adequate representation. It is not evident, however, that such representations are sufficient to support robust face recognition. Depending too much on features, for example, causes problems when the image is degraded by noise or features are occluded (e.g. by sunglasses — see Figure 3-1). We would like to somehow allow for a system to decide what is important to encode for recognition purposes, rather than specifying that initially.

This suggested that an information theory approach of coding and decoding face images may give insight into the information content of face images, emphasizing the significant local and global "features". Such features may or may not be directly related to our intuitive notion of face features such as the eyes, nose, lips, and ears. This may even have important implications for the use of construction tools such as Identikit and Photofit [14], which treat treat faces as "jigsaws" of independent parts.

Such a system motivated by information theory would seek to extract the relevant information in a face image, encode it as efficiently as possible, and compare one face encoding with a database of models encoded similarly. One approach to extracting the information contained in an image of a face is to somehow capture the variation in a collection of face images, independent of any judgement of features, and use this information to encode and compare individual face images.

3.2 Eigenfaces

In mathematical terms, this is equivalent to finding the principal components of the distribution of faces, or the eigenvectors of the covariance matrix of the set of face images, treating an image as a point (or vector) in a very high dimensional space. The eigenvectors are ordered, each one accounting for a different amount of the variation among the face images.

These eigenvectors can be thought of as a set of features which together characterize the variation among face images. Each image contributes some amount to each eigenvector, so that each eigenvector formed from an ensemble of face images appears as a sort of ghostly face image, referred to as an *eigenface*. Examples of these faces are shown in Figure 3-5. Each eigenface deviates from uniform grey where some facial feature differs among the set of training faces; collectively, they map of the variations between faces.

Each individual face image can be represented exactly in terms of a linear combination of the eigenfaces. Each face can also be approximated using only the "best" eigenfaces — those that have the largest eigenvalues, and which therefore account for the most variation within the set of face images. The best M eigenfaces span an M-dimensional subspace — "face space" — of the space of all possible images.

Because eigenfaces will be an orthonormal vector set, the projection of a face image into "face space" is analogous to the well-known Fourier transform. In the FT, an image or signal is projected onto an orthonormal basis set of sinusoids at varying frequencies and phase, as depicted in Figure 3-2(a). Each location of the transformed signal represents the projection onto a particular sinusoid. The original signal or image can be reconstructed exactly by a linear combination of the basis set of signals, weighted by the corresponding component of the transformed signal. If the components of the transform are modified, the reconstruction will be approximate and will correspond to linearly filtering the original signal.

Figure 3-2(b) shows the analogy to the "Eigenface transform". This transform

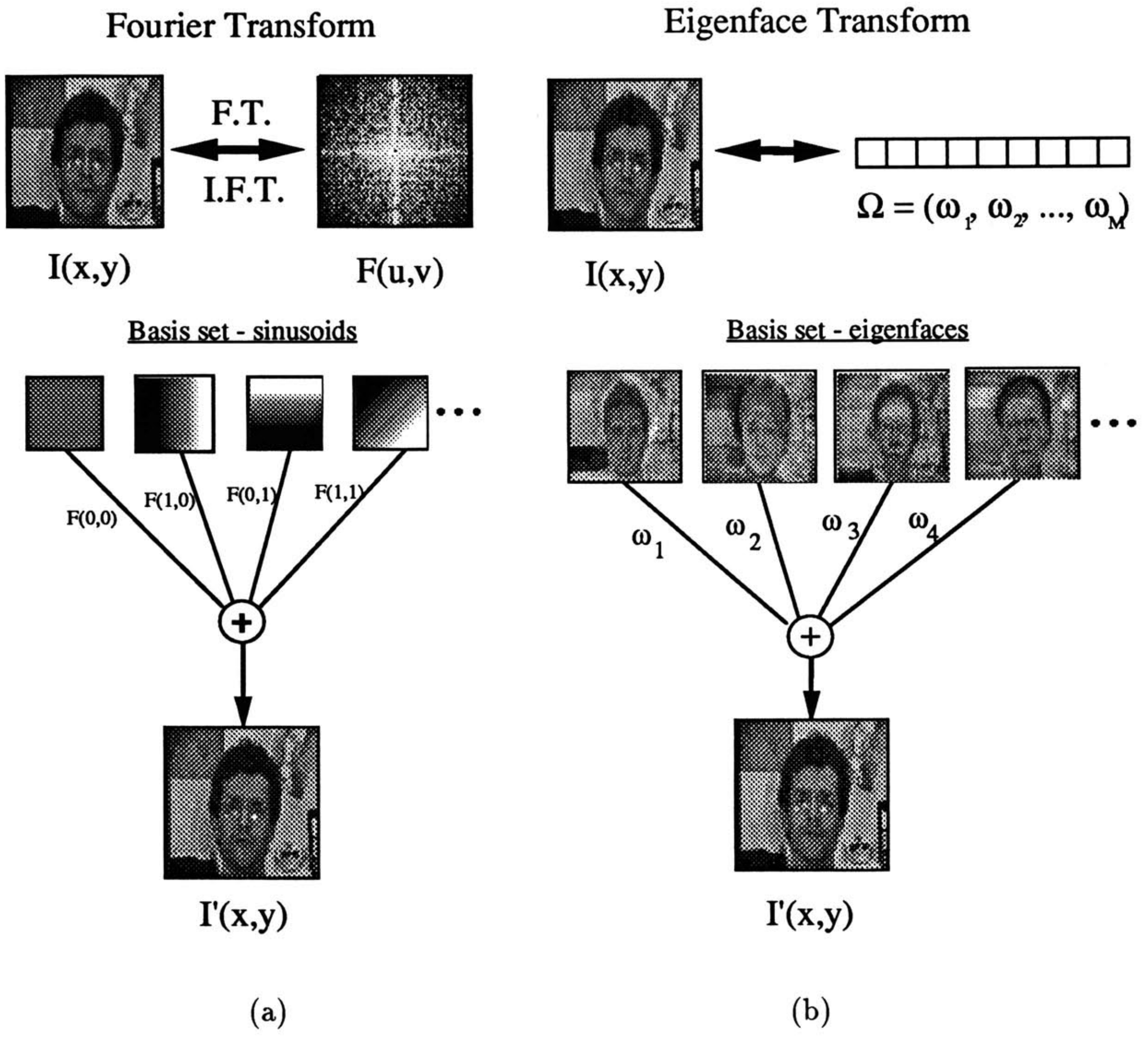


Figure 3-2: Transformation and reconstruction of images with (a) the Fourier transform, and (b) the Eigenface transform.

is non-invertible, in the sense that the basis set is small and can reconstruct only a limited range of images[1]. The transformation will be adequate for recognition to the degree that the "face space" spanned by the eigenfaces can account for a sufficient range of faces.

Principal-component analysis has been applied to pattern recognition tasks for quite some time (e.g. see [54, 57, 21, 103]). Kumar *et al.* [57] proposed a PCA-based filter as optimal for a statistical correlator. Appendix A discusses this work and the relationship between matched filter correlation and recognition using eigenfaces. The idea of using eigenfaces was partially motivated by the work of Sirovich and Kirby [92, 53] for efficiently representing pictures of faces using principal component analysis. Starting with an ensemble of original face images, they calculated a best coordinate system for image compression, where each coordinate is actually an image which they termed an *eigenpicture*. They argued that, at least in principle, any collection of face images can be approximately reconstructed by storing a small collection of weights for each face and a small set of standard pictures (the eigenpictures). The weights describing each face are found by projecting the face image onto each eigenpicture.

Although intended for application to image coding of faces, the eigenpictures do not appear to be sufficient to represent the gamut of facial expressions and viewpoints for the highly accurate reconstruction required in many image coding applications. Sirovich and Kirby's work seemed best suited for applications such as teleconferencing where the accuracy requirements are not as strict and the identity of the speaker is of primary importance.

Face recognition, on the other hand, should not require a precise, low mean-squared-error reconstruction. If a multitude of face images can be reconstructed by weighted sums of a small collection of characteristic features or eigenpictures [92], perhaps an efficient way to learn and recognize faces would be this: build up the characteristic features (eigenfaces) by experience over time and recognize particular faces by comparing the feature weights needed to (approximately) reconstruct them with the weights associated with known individuals. Each individual, therefore, would be characterized by the small set of feature or eigenpicture weights needed to describe and reconstruct them — an extremely compact representation when compared with the images themselves.

[1] For general NxN images, M eigenfaces will span an M-dimensional subspace ("face space") of the huge N^2-dimensional space of all images.

Basing face recognition on this scheme involves an initialization phase where the eigenfaces are constructed from face images, and a continuous processing loop where the eigenfaces are used as a basis for recognition. The one-time initialization operations are:

1. Acquire an initial set of face images (the training set).
2. Calculate the eigenfaces from the training set, keeping only the M eigenfaces which correspond to the highest eigenvalues. These M images define the *face space.*
3. Calculate the corresponding location or distribution in M-dimensional weight space for each known individual, by projecting their face images (from the training set) onto the "face space".

These operations can also be performed occasionally to update or recalculate the eigenfaces as new faces are encountered.

Having initialized the system, the following steps are then used to recognize new face images:

1. Calculate a set of weights based on the input image and the M eigenfaces by projecting the input image onto each of the eigenfaces.
2. Determine if the image is a face at all (whether known or unknown) by checking to see if the image is sufficiently close to "face space" — i.e. determining the ability of the eigenfaces to reconstruct the image.
3. If it is a face, classify the weight pattern as either a known person or as unknown.
4. (Optional) Update the eigenfaces and/or weight patterns.
5. (Optional) If the same unknown face is seen several times, calculate its characteristic weight pattern and incorporate into the known faces, a simple learning mechanism.

The following sections will describe the process in more detail.

3.3 Calculating eigenfaces

Let a face image $I(x, y)$ be a two-dimensional N by N array of (8-bit) intensity values. Such an image may also be considered as a vector of dimension N^2, so that a typical image of size 128 by 128 becomes a vector of dimension 16,384, or, equivalently, a point in 16,384-dimensional space[2]. An ensemble of images maps to a collection of points in this huge space.

Images of faces, being similar in overall configuration, will not be randomly distributed in this huge image space and thus can be described by a relatively low dimensional subspace. The main idea of the principal component analysis (or Karhunen-Loeve expansion) is to find the vectors which best account for the distribution of face images within the entire image space. These vectors define the subspace of face images called "face space". Each vector is of length N^2, describes an N by N image, and is a linear combination of the original face images. Because these vectors are the eigenvectors of the covariance matrix corresponding to the original face images, and because they are face-like in appearance, they are referred to as "eigenfaces."

As a simple example of this analysis, consider "images" of only three pixels. All possible 1x3 images fill a three-dimensional space[3]. An image of this type is fully specified by three numbers, its coordinates in the 3-D space in Figure 3-3(a). If a collection of these images occupy a two-dimensional subspace as in Figure 3-3(b), they can be exactly specified by just two numbers, the projections onto the vectors $\mathbf{u_1}$ and $\mathbf{u_2}$ which describe the plane (span the subspace). These vectors are the significant eigenvectors of the covariance matrix of the images. Because they are vectors in the 3-D space, they can also be "displayed" as three-pixel images. A new image which lies near the 2D plane can now be approximately represented by its projection into the plane (or equivalently its projection onto the eigenvectors).

This example is directly analogous to the construction and use of eigenfaces. With real images, the original space has dimension much greater than three, e.g. 16,384-dimensional for 128 by 128 images. The important assumption (supported by [92]) is that a collection of face images spans some low-dimensional subspace, similar to the plane of points in the example. The eigenvectors (eigenfaces in this case) are

[2] The analysis is equivalent for non-square images.

[3] We will for now ignore the quantization and limited range of the space determined by the limited precision discrete pixel values.

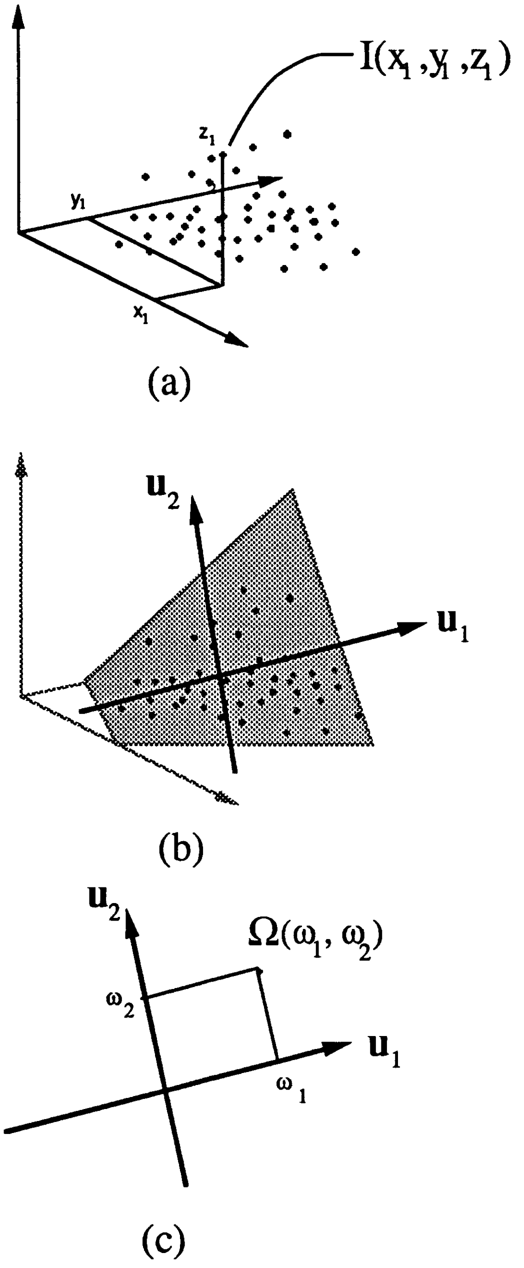


Figure 3-3: Simple example of principal component analysis. (a) Images with three pixels are described as points in three-space. (b) The subspace defined by a planar collection of these images is spanned by two vectors. One choice for this pair of vectors is the eigenvectors of the covariance matrix of the ensemble, $\mathbf{u_1}$ and $\mathbf{u_2}$. (c) Two coordinates are now sufficient to describe the points, or images: their projections onto the eigenvectors, (ω_1, ω_2).

16,384-dimensional, and may be viewed as images. As we will see in later sections, there are two important measurements when evaluating a new image: (1) its distance *away from* the subspace (face space) spanned by the eigenfaces, and (2) the position of its projection *into* the face space relative to known faces.

Let the training set of face images be $\phi_1, \phi_2, \phi_3, ...\phi_M$. The average face of the set is defined by $\Psi = \frac{1}{M}\sum_{n=1}^{M} \phi_n$. Each face differs from the average by the vector $\Phi_i = \phi_i - \Psi$. An example training set is shown in Figure 3-4(a), with the average face Ψ shown in Figure 3-4(b)[4]. This set of very large vectors is then subject to principal component analysis, which seeks a set of (M-1) orthonormal vectors, $\mathbf{u}_n$, which best describes the distribution of the data. The kth vector, $\mathbf{u}_k$, is chosen such that

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(\mathbf{u}_k^t \Phi_n)^2 \tag{3.1}$$

is a maximum, subject to

$$\mathbf{u}_l^t \mathbf{u}_k = \delta_{lk} = \begin{cases} 1, & \text{if } l = k \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

for $l < k$, which constrains the vectors to be orthogonal.

The vectors $\mathbf{u}_k$ and scalars λ_k are the significant M eigenvectors and eigenvalues, respectively, of the covariance matrix

$$\begin{aligned} C &= \tfrac{1}{M}\textstyle\sum_{n=1}^{M} \Phi_n \Phi_n^t \\ &= AA^t \end{aligned} \tag{3.3}$$

where the matrix $A = [\ \Phi_1\ \Phi_2\ ...\ \Phi_M\ ]$. The matrix C, however, is N^2 by N^2, and determining the N^2 eigenvectors and eigenvalues is an intractable task for typical image sizes. We need a computationally feasible method to find these eigenvectors $\mathbf{u}_i$ of C:

$$AA^t\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{3.4}$$

If the number of data points in the image space is less than the dimension of the space ($M < N^2$), there will be only $M - 1$, rather than N^2, meaningful eigenvectors.

[4] Elimination of the background will be discussed later, but for now the face and background are not distinguished.

(The remaining eigenvectors will have associated eigenvalues of zero.) Fortunately we can solve for the N^2-dimensional eigenvectors in this case by first solving for the eigenvectors of an M by M matrix — e.g. solving a 16x16 matrix rather than a 16,384 by 16,384 matrix — and then taking appropriate linear combinations of the face images Φ_i. Consider the eigenvectors $\mathbf{v}_i$ of A^tA such that

$$A^tA\mathbf{v}_i = \mu_i\mathbf{v}_i. \tag{3.5}$$

Premultiplying both sides by A, we have [93]

$$AA^tA\mathbf{v}_i = \mu_iA\mathbf{v}_i \tag{3.6}$$

or

$$AA^t(A\mathbf{v}_i) = \mu_i(A\mathbf{v}_i) \tag{3.7}$$

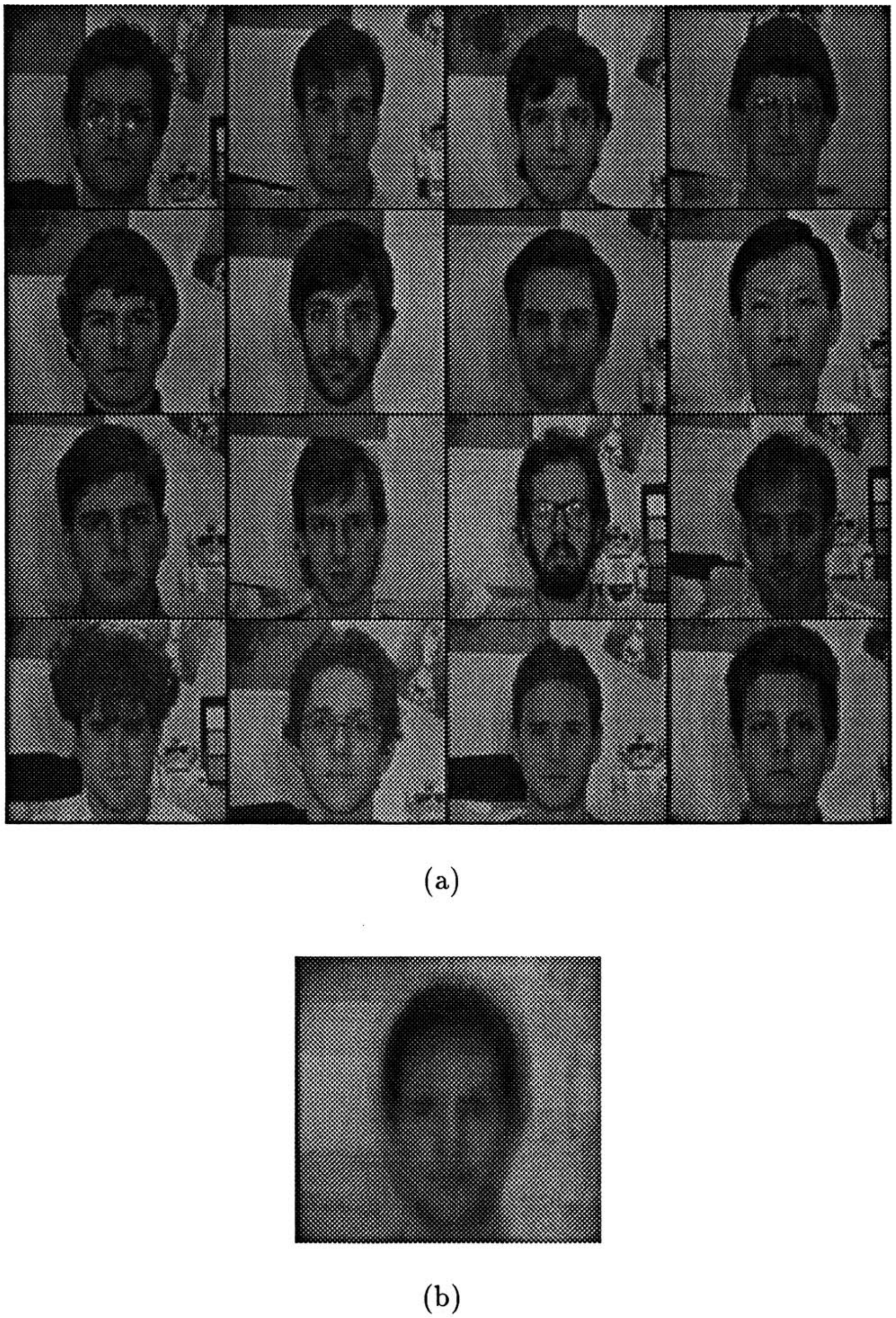
and comparing with Equation 3.4 we see that $A\mathbf{v}_i$ are the eigenvectors of $C = AA^t$.

Following this analysis, we construct the M by M matrix $L = A^tA$, where $L_{mn} = \Phi_m^t\Phi_n$, and find the M eigenvectors, $\mathbf{v}_l$, of L. These vectors determine linear combinations of the M training set face images to form the eigenfaces $\mathbf{u}_l$:

$$\mathbf{u}_l = A\mathbf{v}_i \tag{3.8}$$

With this analysis the calculations are greatly reduced, from the order of the number of pixels in the images (N^2) to the order of the number of images in the training set (M). In practice, the training set of face images will be relatively small ($M \ll N^2$), and the calculations become quite manageable. The associated eigenvalues allow us to rank the eigenvectors according to their usefulness in characterizing the variation among the images, and therefore to choose a significant subset to keep. Figure 3-5 shows the top seven eigenfaces derived from the input images of Figure 3-4.

The issue of choosing how many eigenfaces to keep for recognition involves a tradeoff between recognition accuracy and processing time. Each additional eigenface adds to the computation involved in classifying and locating a face. This is not vital for small databases, but as the size of the database increases it becomes relevant. In the examples of this chapter and the experiments described in the next chapter, a



(a)

(b)

Figure 3-4: (a) Face images used as the training set, including the background. (b) The average face Ψ.

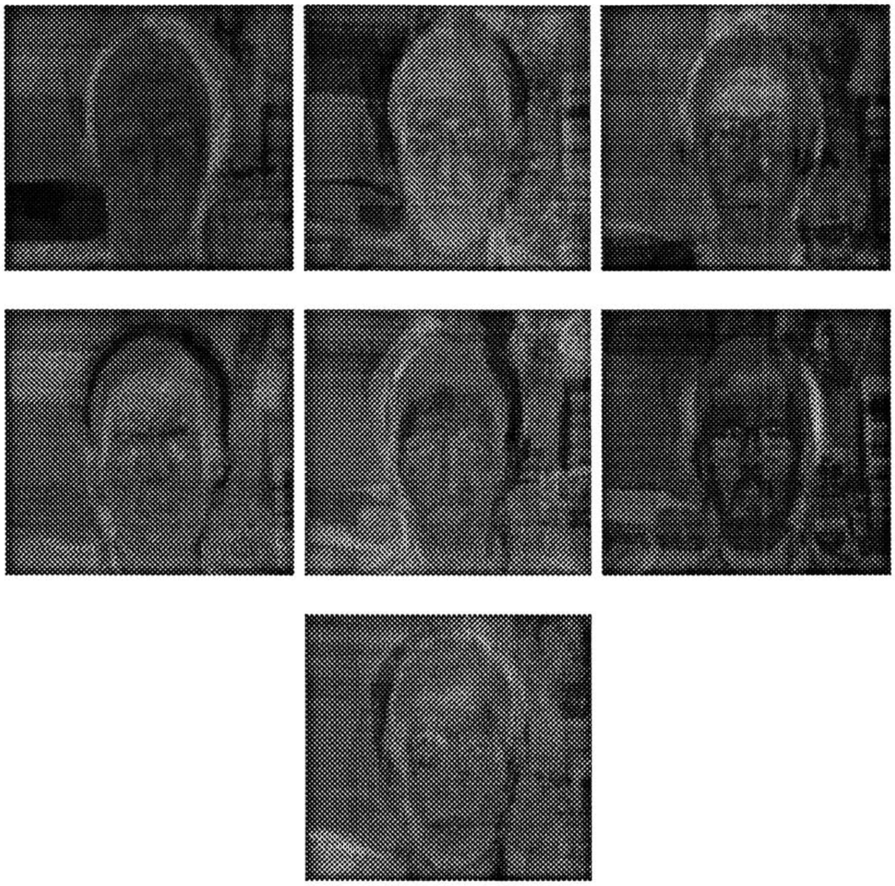


Figure 3-5: Seven of the eigenfaces calculated from the input images of Figure 3-4, without the background removed. (See Section 3.6.1 regarding the background.).

heuristic evaluation chose seven or eight eigenfaces to use from a database of sixteen face images.

3.4 Using eigenfaces to classify a face image

The eigenface images calculated from the eigenvectors of the matrix L span a basis set with which to describe face images. Sirovich and Kirby [92] evaluated a limited version of this framework on an ensemble of $M = 115$ images of caucasian males, digitized in a controlled manner, and found that about 40 eigenfaces were sufficient for a very good description of the set of face images. Using $M' = 40$ eigenfaces, RMS pixel-by-pixel errors in representing cropped versions of face images were about 2%.

Since eigenfaces seem adequate for describing face images under very controlled

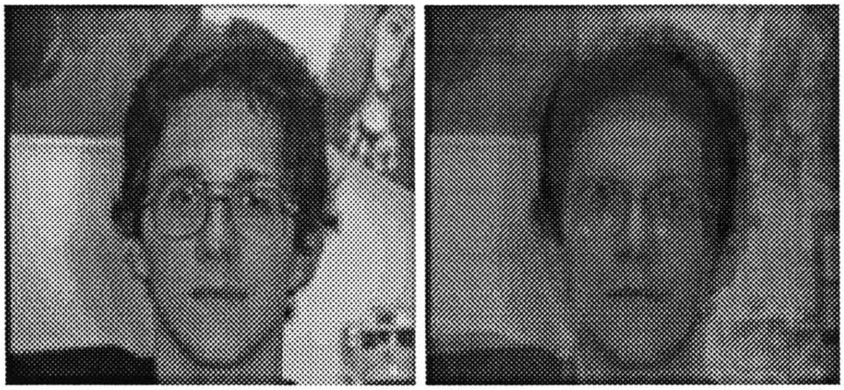


Figure 3-6: An original face image and its projection onto the face space defined by the eigenfaces of Figure 3-5.

conditions, it was decided to investigate their usefulness as a tool for face identification. In practice, a smaller M' is sufficient for identification, since accurate reconstruction of the image is not a requirement. In this framework, identification becomes a pattern recognition task. The eigenfaces span an M'-dimensional subspace of the original N^2 image space. The M' significant eigenvectors of the L matrix are chosen as those with the largest associated eigenvalues. As previously mentioned, in many of these test cases, $M' = 7$ eigenfaces were used from $M = 16$ face images.

A new face image (ϕ) is transformed into its eigenface components ω_i (projected into "face space") by a simple operation,

$$\omega_k = \mathbf{u}_k^t(\phi - \Psi), \tag{3.9}$$

for $k = 1, \ldots, M'$. The average face Ψ is subtracted and the remainder is projected onto the eigenfaces $\mathbf{u}_k$. This describes a set of point-by-point image multiplications and summations, operations performed at approximately frame rate on current image processing hardware. Figure 3-6 shows an image and its projection into the (in this case) seven-dimensional face space.

The weights form a vector $\Omega^t = [\omega_1 \; \omega_2 \; \ldots \; \omega_{M'}]$ that describes the contribution of each eigenface in representing the input face image, treating the eigenfaces as a basis set for face images. The vector is then used in a standard pattern recognition algorithm to find which of a number of pre-defined face classes, if any, best describes the face. The simplest method for determining which face class provides the best description of an input face image is to find the face class k that minimizes the

Euclidian distance

$$\epsilon_k = \|(\Omega - \Omega_k)\|^2, \tag{3.10}$$

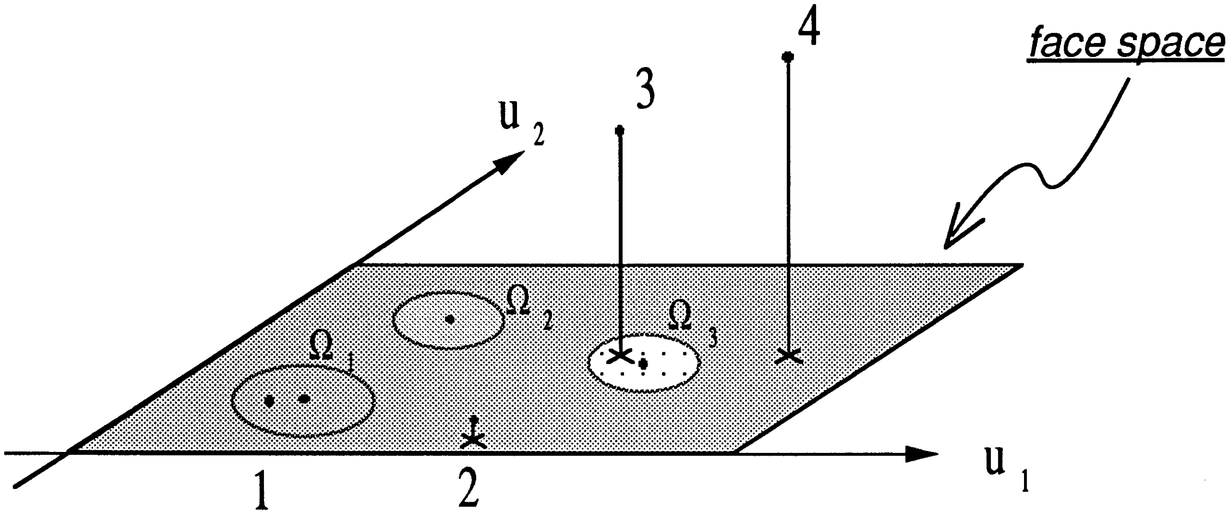
where Ω_k is a vector describing the kth face class. The face classes Ω_i are calculated by averaging the results of the eigenface representation over a small number of face images (as few as one) of each individual. A face is classified as belonging to class k when the minimum ϵ_k is below some threshold θ_ϵ. Otherwise the face is classified as "unknown", and optionally used to create a new face class.

The nearest-neighbor classification assumes a uniform gaussian distribution in face space of an individual's feature vectors Ω_i. Since there is no *a priori* reason to assume such a distribution, we want to characterize it rather than assume it is gaussian. The class distribution can be obtained over a short time by continuously projecting the images of an individual onto the eigenfaces, keeping track of the projection values while allowing for variations in the subject's expression, the lighting, etc. The data is then fit to a non-uniform multidimensional gaussian which describes an individual's distribution in face space. This has been tested but not yet implemented into the working recognition system. Non-linear networks such as described by Fleming and Cottrell [36] appear to be a promising way to learn more complex face space distributions by example.

Because creating the vector of weights is equivalent to projecting the original face image onto the low-dimensional face space, many images (most of them looking nothing like a face) will project onto a given pattern vector. In many pattern recognition schemes this will be a false positive, incorrectly identified as a match. This is not a problem for the system, however, since the distance ε between the image and the face space gives a direct measure of the "faceness", or how well the eigenfaces describe the image. This is simply the squared distance between the mean-adjusted input image $\Phi = \phi - \Psi$ and $\Phi_f = \sum_{i=1}^{i=M'} \omega_k \mathbf{u}_k$, its projection onto face space:

$$\varepsilon^2 = \|\Phi - \Phi_f\|^2 \tag{3.11}$$

If this distance ε — the distance from face space — is large, the image is not well described by the eigenfaces and therefore is not considered a face. A face image, on the other hand, should lie near the face space, and so produce a small ε. We choose a threshold β_ε to represent the minimum acceptable distance from face space.



	Face space	Known face class	Result
1	near	near	Recognized as Ω_1
2	near	far	Who are you?
3	far	near	?False positive?
4	far	far	No face

Figure 3-7: A simplified version of face space to illustrate the four results of projecting an image into face space. In this case, there are two eigenfaces ($\mathbf{u}_1$ and $\mathbf{u}_2$) and three known individuals (Ω_1, Ω_2, and Ω_3).

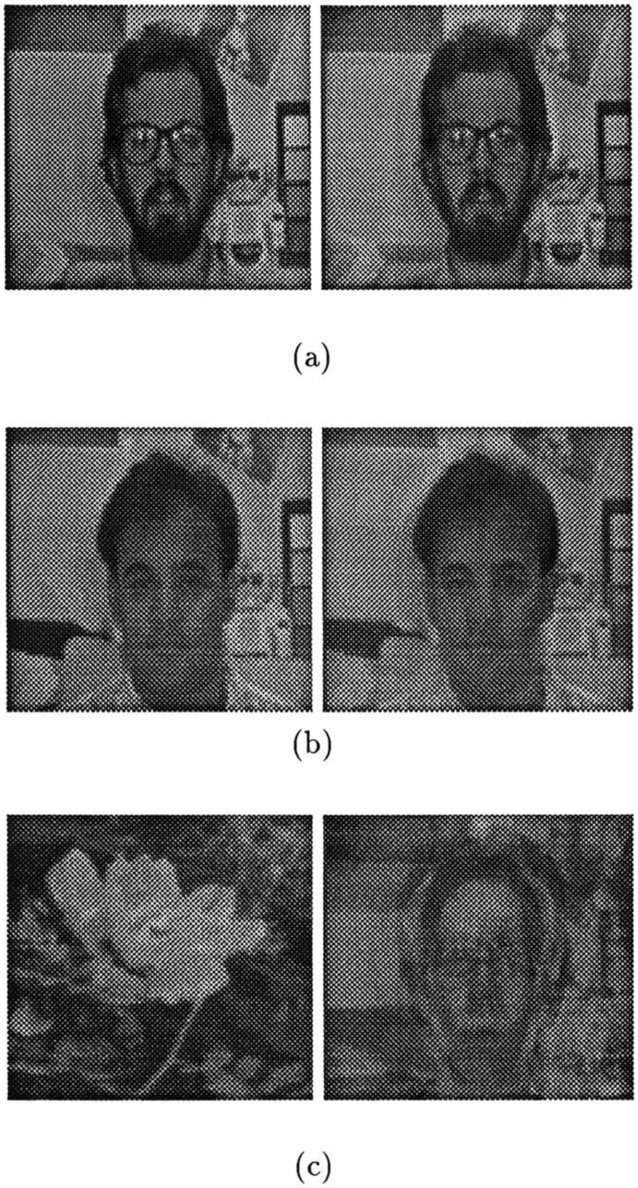
Thus there are four possibilities for an input image and its pattern vector: (1) near face space and near a face class; (2) near face space but not near a known face class; (3) distant from face space and near a face class; and (4) distant from face space and not near a known face class. "Near" and "distant" are defined relative to the threshold values β_ϵ and θ_ϵ.

Figure 3-7 shows a simple example of these cases, with two eigenfaces ($\mathbf{u}_1$ and $\mathbf{u}_2$) and three known individuals (face classes Ω_1, Ω_2, and Ω_3). In the first case, an individual is recognized and identified as person 1 because it is very close to the corresponding face class Ω_1). In the second case, an unknown individual is present, since the image is "face-like" (near face space), but not close to any of the known face classes. The last two cases indicate that the image is not of a face. Case three typically shows up as a false positive in most recognition systems; in our framework, however, the false recognition may be detected because of the significant distance from face space (large ε). Figure 3-8 shows some images and their projections into face space and gives a measure of distance from the face space for each.

3.5 Using eigenfaces to detect and locate faces

The analysis in the preceding sections assumes we have a centered face image. We need some way, then, to locate and center a face in a scene in order to do the recognition. The idea of projecting an image into face space (equivalent to reconstructing the image using the eigenfaces) and finding the distance ε between the original and the reconstruction is useful here, as it gives a measure of "faceness" for every subimage in a larger scene.

As seen in Figure 3-8, images of faces do not change radically when projected into the face space, while the projection of non-face images appear quite different. This basic idea is used to detect the presence of faces in a scene: at every location in the image, calculate the distance ε between the local subimage and its face space projection. This *distance from face space* is used as a measure of "faceness", so the result of calculating the distance from face space at every point in the image is a "face map" $\varepsilon(x, y)$. Figure 3-9 shows an image and its face map — low values (the dark area) indicate the presence of a face. Local minima in the face map indicate possible faces; if the value of $\varepsilon(x, y)$ at any minima is below a threshold, a face is detected. In Figure 3-9(b), the distinct minimum is correctly located in the center of the face.

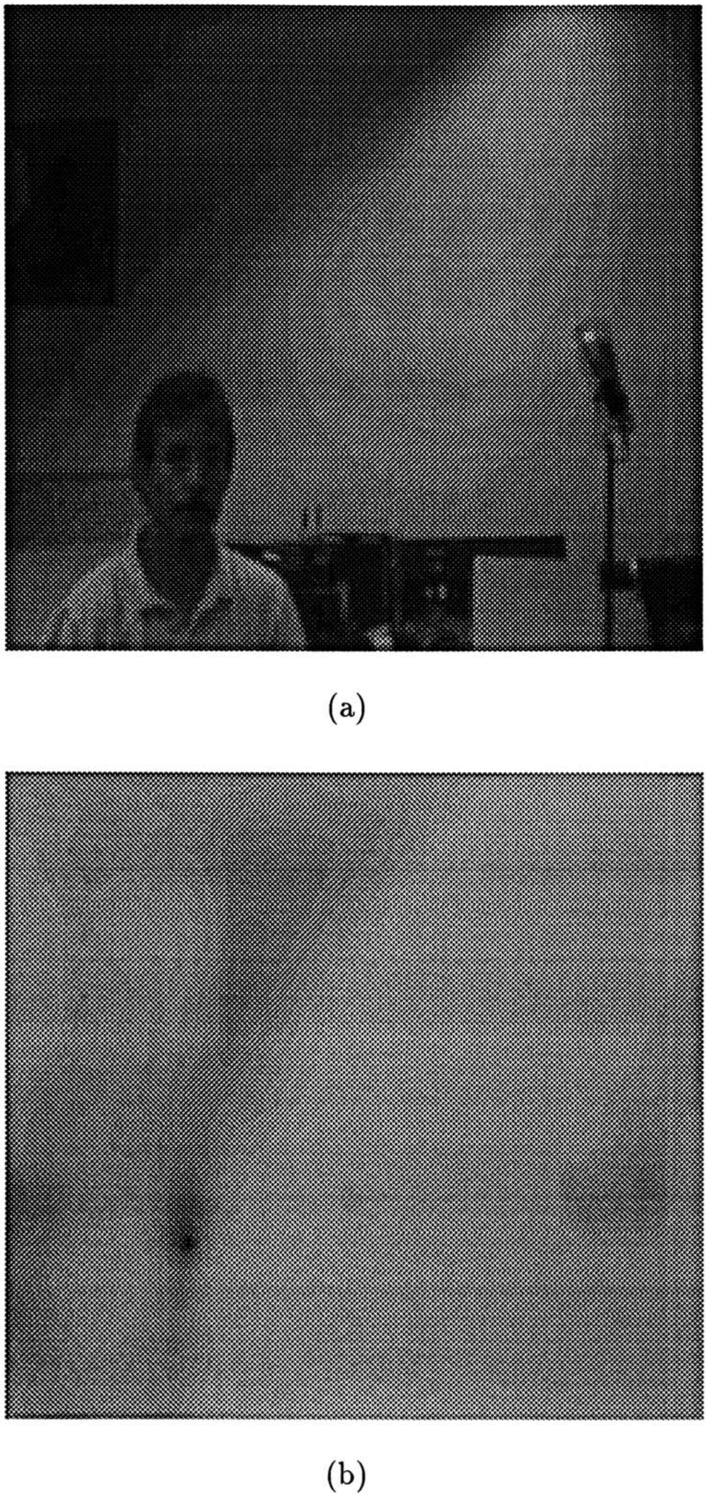


(a)

(b)

(c)

Figure 3-8: Three images and their projections onto the face space defined by the eigenfaces of Figure 3-5. The relative measures of distance from face space (ε) are: (a) 29.8 (b) 58.5 (c) 5217.4. Images (a) and (b) are in the original training set.



(a)

(b)

Figure 3-9: (a) Original image. (b) Face map, where low values (dark areas) indicate the presence of a face.

Unfortunately, direct application of Equation 3.11 at every pixel is rather expensive. By manipulating the equation and implementing part of the computation via the fast fourier transform, we can produce an efficient method of calculating the face map $\varepsilon(x, y)$ in the following manner.

To calculate the face map at every pixel of an image $I(x, y)$, we need to project the subimage centered at that pixel onto face space, then subtract the projection from the original. To project a subimage ϕ onto face space, we must first subtract the mean image, resulting in $\Phi = \phi - \Psi$. With Φ_f being the projection of Φ onto face space, the distance measure at a given image location is then:

$$\begin{aligned} \varepsilon^2 &= ||\Phi - \Phi_f||^2 \\ &= (\Phi - \Phi_f)^t(\Phi - \Phi_f) \\ &= \Phi^t\Phi - \Phi^t\Phi_f - \Phi_f^t(\Phi - \Phi_f) \\ &= \Phi^t\Phi - \Phi^t\Phi_f \end{aligned} \tag{3.12}$$

since $\Phi_f \perp (\Phi - \Phi_f)$ and

$$\varepsilon^2 = \Phi^t\Phi - \Phi_f^t\Phi_f \tag{3.13}$$

since $\Phi^t\Phi_f = \Phi_f^t\Phi_f$. Because Φ_f is a linear combination of the eigenfaces ($\Phi_f = \sum_{i=1}^{M} \omega_i \mathbf{u}_i$) and the eigenfaces are orthonormal vectors, we have

$$\begin{aligned} \Phi_f^t\Phi_f &= (\omega_1\mathbf{u}_1 + \omega_2\mathbf{u}_2 + \ldots)^t(\omega_1\mathbf{u}_1 + \omega_2\mathbf{u}_2 + \ldots) \\ &= (\omega_1\mathbf{u}_1^t\omega_1\mathbf{u}_1 + \omega_1\mathbf{u}_1^t\omega_2\mathbf{u}_2 + \ldots + \omega_2\mathbf{u}_2^t\omega_2\mathbf{u}_2 + \ldots) \\ &= \textstyle\sum_{i=1}^{M} \omega_i^2 \end{aligned} \tag{3.14}$$

Therefore

$$\varepsilon^2 = \Phi^t\Phi - \sum_{i=1}^{M} \omega_i^2 \tag{3.15}$$

at every pixel location in the image, or

$$\varepsilon^2(x, y) = \Phi^t(x, y)\Phi(x, y) - \sum_{i=1}^{M} \omega_i^2(x, y) \tag{3.16}$$

where $\varepsilon(x, y)$ and $\omega_i(x, y)$ are scalar functions of image location, and $\Phi(x, y)$ is a vector function of image location.

The second term of Equation 3.16 is calculated in practice by a correlation with the L eigenfaces:

$$\begin{aligned} \sum_{i=1}^{M} \omega_i^2(x,y) &= \sum_{i=1}^{M} \Phi^t(x,y)\mathbf{u}_i \\ &= \sum_{i=1}^{M} (\phi(x,y) - \Psi)^t \mathbf{u}_i \\ &= \sum_{i=1}^{M} (\phi^t(x,y)\mathbf{u}_i \; - \; \Psi^t \mathbf{u}_i) \\ &= \sum_{i=1}^{M} (I(x,y) \otimes \mathbf{u}_i \; - \; \Psi^t \mathbf{u}_i) \end{aligned} \tag{3.17}$$

where $\otimes$ is the correlation operator and $I(x,y)$ is the original image. The first term of Equation 3.16 becomes

$$\begin{aligned} \Phi^t(x,y)\Phi(x,y) &= (\phi(x,y) - \Psi)^t(\phi(x,y) - \Psi) \\ &= \phi^t(x,y)\phi(x,y) - 2\phi^t(x,y)\Psi + \Psi^t\Psi \\ &= \phi^t(x,y)\phi(x,y) - 2I(x,y) \otimes \Psi + \Psi^t\Psi \end{aligned} \tag{3.18}$$

so that

$$\begin{aligned} \varepsilon^2(x,y) = \phi^t(x,y)\phi(x,y) - 2I(x,y) \otimes \Psi + \Psi^t\Psi \\ - \sum_{i=1}^{M} (I(x,y) \otimes \mathbf{u}_i \; - \; \Psi^t \mathbf{u}_i) \end{aligned} \tag{3.19}$$

Since the average face Ψ and the eigenfaces $\mathbf{u}_i$ are fixed, Equation 3.19 becomes

$$\varepsilon^2(x,y) = \phi^t(x,y)\phi(x,y) - 2I(x,y) \otimes \Psi - \sum_{i=1}^{M} I(x,y) \otimes \mathbf{u}_i \; + \; C \tag{3.20}$$

where the constant $C = \Psi^t\Psi + \sum_{i=1}^{M} \Psi^t \mathbf{u}_i$ may be computed only once before the recognition process begins.

Thus the computation of the face map involves only $M + 1$ correlations over the input image and the computation of the first term $\phi^t(x,y)\phi(x,y)$. This is computed by squaring the input image $I(x,y)$ and, at each image location, summing the squared values of the local subimage by convolving the result with a mask of all 1's. (The effect of the background is eliminated by modifying this mask to be 1 in the face area and 0 elsewhere — this is the binary mask described in Section 3.6.1).

The correlations are implemented in the face recognition system as a series of FFTs, while the remaining operations are simple addition and point operations. Timing is discussed in Chapter 5. As discussed in Chapter 6, these computations can be implemented by a simple neural network.

3.6 Recognition issues

The preceding sections describe the basic eigenfaces formulation. A number of other issues must be addressed to achieve a robust working system. In this section I discuss some of these issues and indicate current or proposed solutions.

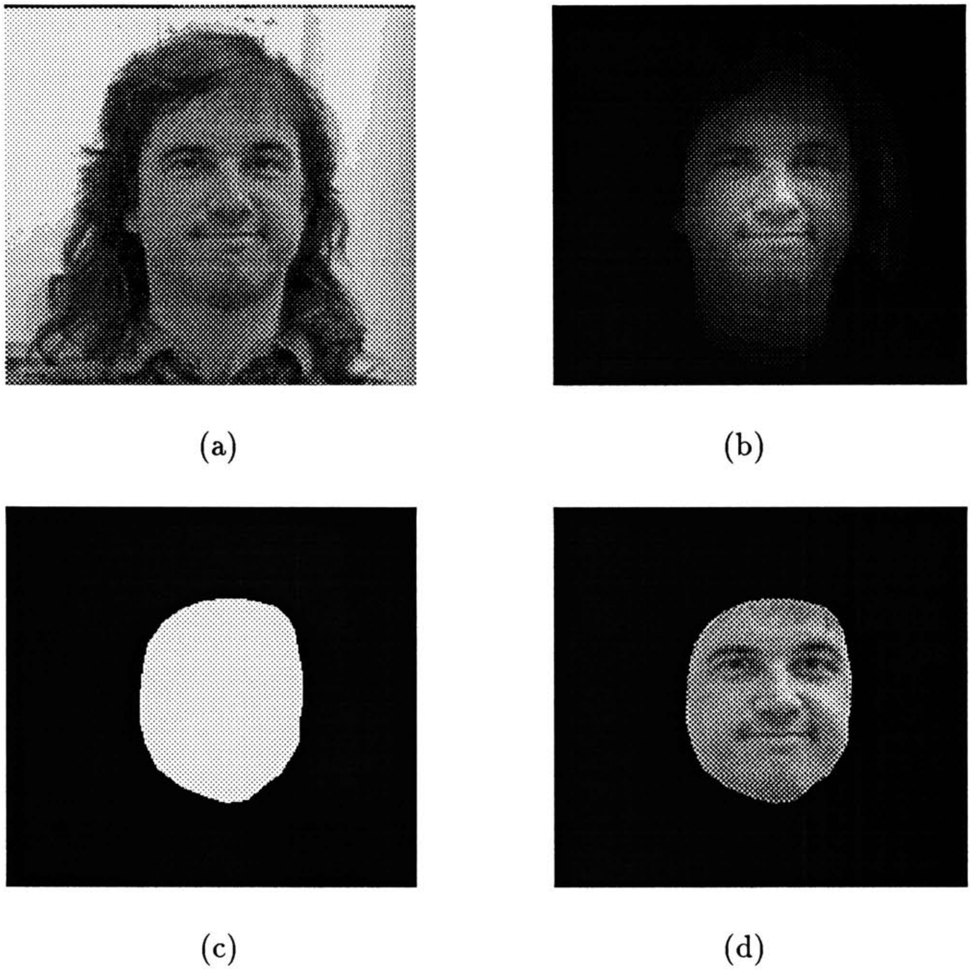
3.6.1 Eliminating the background

In the preceding analysis we have ignored the effect of the background. In practice, the background can significantly affect the recognition performance, since the eigenface analysis as described above does not distinguish the face from the rest of the image. In the experiments described in Chapter 4, it was necessary to reduce or eliminate the background from the database of face images.

I have used two methods in order to handle this problem without having to solve other difficult vision problems such as robust segmentation of the head. The first consists of multiplying the input face image by a two-dimensional non-uniform gaussian window centered on the face, as in Figure 3-10(b), thus diminishing the background and accentuating the middle of the face. Experiments in human strategies of face recognition [48] cite the importance of the internal facial features for recognition of familiar faces. De-emphasizing the outside of the face is also a practical consideration since changing hairstyles may otherwise negatively affect the recognition. This technique was moderately effective in reducing background effects, but at the expense of complicating the "distance from face space" measurement.

A more useful and efficient technique is to remove the background from the very beginning, when grabbing the training set of face images. This is done simply by the operator outlining the head or face of the first training set image, using a mouse or digitizing tablet. From this outline, a binary mask is made defining the face region for all subsequent processing, and each face image in the training set is multiplied by this face mask, as shown in Figure 3-10(b) and (c). The background of the training set images is therefore consistently zero. Because the eigenfaces are made from a training set with the background masked out of each image, they will also have values of zero at these image locations.

Because the background is zero in the eigenfaces it does not contribute at all to the projection of a new image into face space or to the subsequent classification. The



(a)

(b)

(c)

(d)

Figure 3-10: Two methods to reduce or eliminate the effect of background. (a) An original face image. (b) Multiplied by a gaussian window, emphasizing the center of the face. (c) The binary face mask outlined by the operator (while gathering the training set). (d) The resulting database entry.

"distance from face space" measurement is modified only by changing a mask of all ones to be zero in the background as mentioned in the previous Section. The computation is not increased at all. This method has proven quite effective in eliminating background problems.

3.6.2 Scale (head size) and orientation invariance

The experiments of Chapter 4 will show that recognition performance decreases quickly as the head size, or scale, is misjudged. The head size in the input image must be close to that of the eigenfaces for the system to work well. The motion analysis which will be discussed in Chapter 5 can give an estimate of head width, from which the face image is rescaled to match the eigenface size. However the current system is limited to one moving person with no camera motion.

Another approach to the scale problem, which may be separate from or in addition to the motion estimate, is to use multiscale eigenfaces, in which an input face image is compared with eigenfaces initially at a number of scales. In this case the image will appear to be near the face space of only the closest scale eigenfaces. Equivalently, we can scale the input image to multiple sizes and choose the scale which results in the smallest distance measure to face space using a single set of eigenfaces. A two-pass strategy is to first look for the best match (the smallest $\varepsilon(x, y)$) with eigenfaces spaced an octave apart, and then refine the search around the best octave using eigenfaces of different scale within an octave.

Although the eigenfaces approach is not extremely sensitive to 2-D head orientation (i.e. sideways tilt of the head), a non-upright view will cause some performance degradation. An accurate estimate of the head tilt will certainly benefit the recognition. Two simple methods have been considered and tested for estimating head orientation. The first is to calculate the orientation of the motion blob of the head. This is less reliable as the shape tends toward a circle, however. Using the fact that faces are reasonably symmetric patterns, at least for frontal views, I have tested simple symmetry operators to estimate head orientation. Once the orientation is estimated, the image can be rotated to align the head with the eigenfaces. A more sophisticated symmetry operator, demonstrated in real time on faces, is described by Reisfeld *et al.* [78].

3.6.3 Multiple views

For most applications, a face recognition system needs to deal with the range of facial views from frontal to profile. This range can be incorporated into the eigenface approach by defining a limited number of face classes for each known person corresponding to *characteristic views.* For example, an individual may be represented by face classes corresponding to a frontal face view, oblique views at ± 45°, and right and left profile views. In many interactive viewing situations a small number of views will be sufficient to recognize a face. This is partly because people are likely to move into a position close to one of these characteristic views, and partly because of the associative memory nature of the approach — each set of eigenfaces can deal with a small range of viewpoints rather than one exact view.

To represent multiple face classes, we need multiple sets of eigenfaces, one set for each view. The first step of recognition is to calculate the distance to each separate face space (or "view space"), resulting in a number of distance maps $\varepsilon_k(x, y)$. Next we select the view with the minimum $\varepsilon(x, y)$, and proceed with the recognition within the chosen view.

A more efficient method is to lump images from all the views into one set of eigenfaces. This has the advantage of using fewer than $3M$ eigenfaces, since there will be at least some correlation among the images of different views. (The more highly correlated the training set of images, the fewer eigenfaces are needed.) So it can be faster and simpler to set up. A potential disadvantage is that many images will appear to be close to the face space which are not normal face images — e.g. linear combinations of different views. Because these are unlikely to appear in real situations, however, this is not a significant problem. Another complication is that of solving an order $3M$ eigenvalue problem rather than 3 such problems of order M. For reasonably small databases this is also not a significant problem. However the background is harder to eliminate, since the common mask described in Section 3.6.1 must be large enough to include disparate views.

Tests of the system using three views are promising. Figure 3-11 shows a simple example of the recognition of three views of two people. Testing with multiple views of many people is underway.

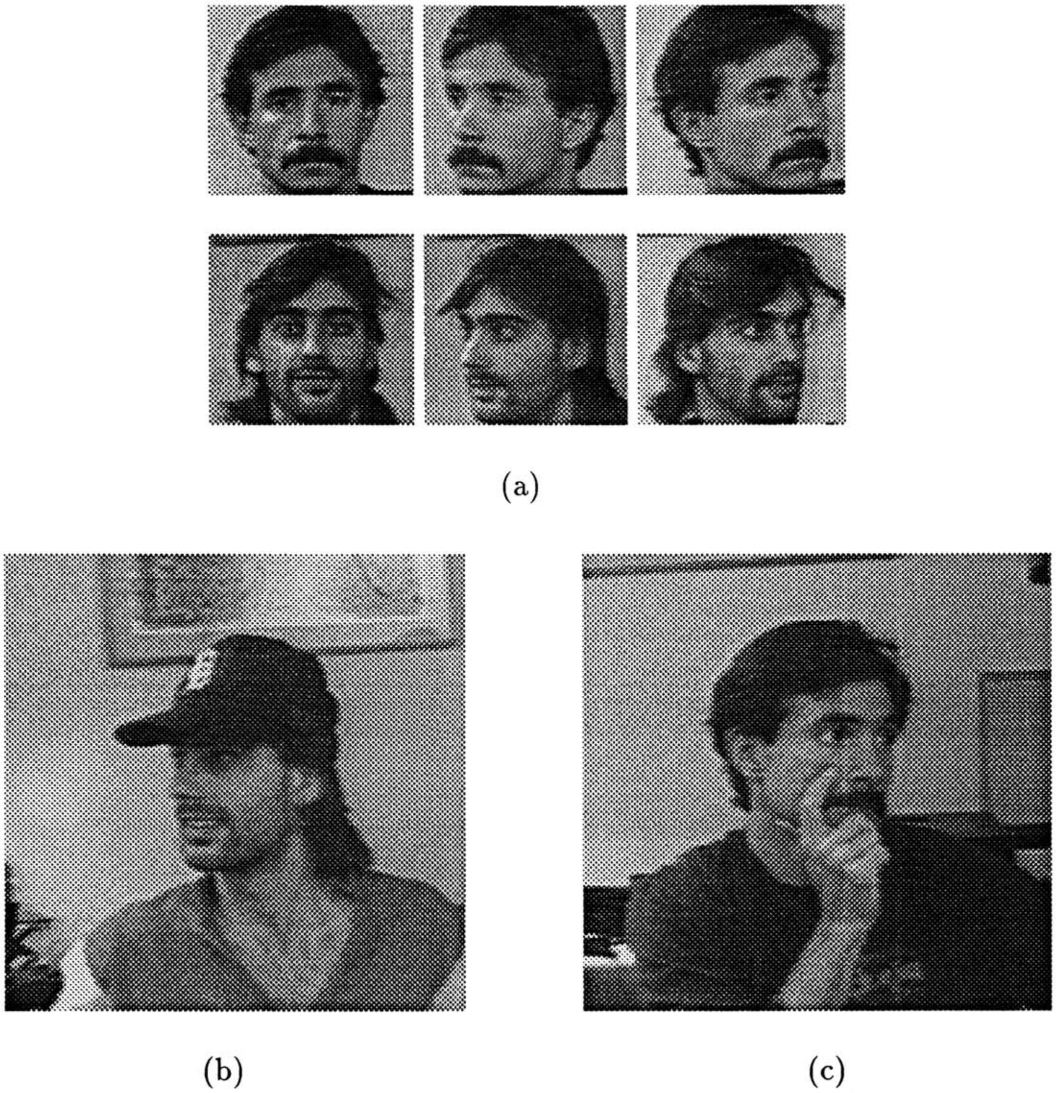


Figure 3-11: A simple multiple-view recognition example: two people, three views. (a) The training set. (b) Image correctly recognized as the left view of person 2. (c) Image correctly recognized as the right view of person 1.

3.6.4 Learning to recognize new faces

The idea of projecting into face space creates the ability to learn and subsequently recognize new faces in an unsupervised manner. When an image is sufficiently close to face space but is not classified as one of the familiar faces (case 2 in Figure 3-7), it is initially labeled as "unknown". The system stores the pattern vector and possibly the corresponding unclassified image. If a collection of "unknown" pattern vectors cluster together in the pattern space, the presence of a new but unidentified face is postulated.

Depending on the application, the system may then alert the user (e.g. a security guard) that an unknown person is present, or continue on to learn the new face for subsequent recognition. The images corresponding to the pattern vectors in the cluster are then checked for similarity by requiring that the distance from each image to the mean of the images is less than a predefined threshold. If the images pass the similarity test, the average of the feature vectors is added to the face classes as a new person. Also, the supposed face image closest to the mean image is added to the database of known faces. Occasionally, the eigenfaces may be recalculated using these stored images as part of the new training set. When this is done, the system has effectively learned the new face.

Figure 3-12 illustrates learning a new face. A training set of faces (front views only) gives rise to the eigenfaces of Figure 3-12(a). A new person, not in the database, enters the scene, and the motion processing (described in Chapter 5 and the distance from face space measure locate the face as shown in Figure 3-12(b) and (c). Because this is reasonably close to face space but not close to any of the known classes, the face is considered as a new person and added to the database.

3.7 Summary of the recognition procedure

To summarize, the eigenfaces approach to face recognition involves the following steps:

1. Collect a set of characteristic face images of the known individuals. This set should include a number of images for each person, with some variation in expression and in the lighting. (Say four images of ten people, so $M = 40$.)

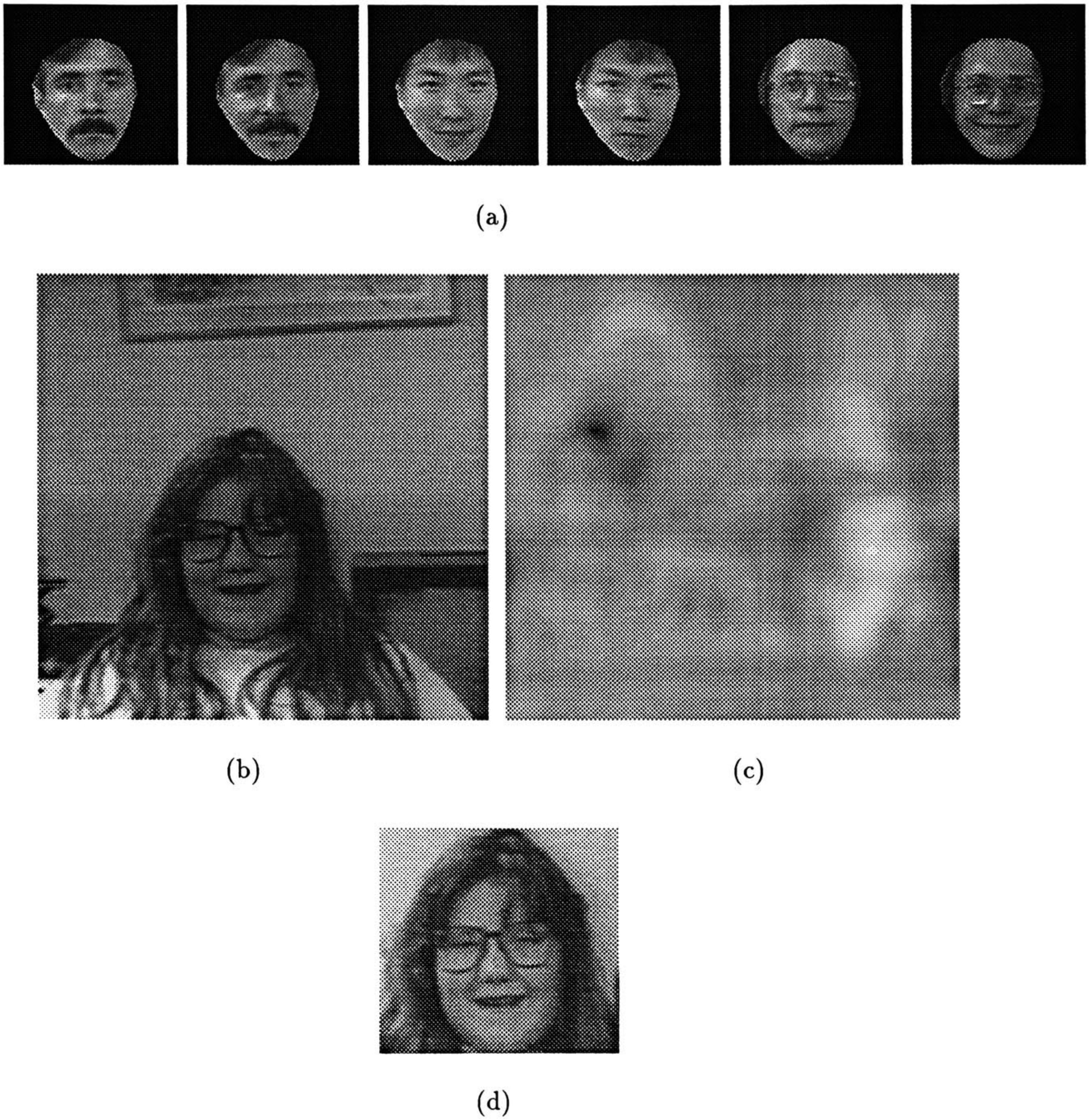


Figure 3-12: An example of learning a new face. (a) The training set of 3 people. (b) An unknown face in the scene, located coursely via motion processing. (c) Distance from face space map $\varepsilon(x, y)$ for the scene. (d) The new face, located at the minimum $\varepsilon(x, y)$. This face image was close enough to face space to be considered possibly a face, but did not project near a known class in face space.

2. Calculate the (40x40) matrix L, find its eigenvectors and eigenvalues, and choose the M' eigenvectors with the highest associated eigenvalues. (Let $M' = 10$ in this example.)

3. Combine the normalized training set of images according to Equation 3.8 to produce the ($M' = 10$) eigenfaces $\mathbf{u}_k$.

4. For each known individual, calculate the class vector Ω_k by averaging the eigenface pattern vectors Ω (from Equation 3.10) calculated from the original (four) images of the individual. Choose a threshold θ_ϵ which defines the maximum allowable distance from any face class, and a threshold β_ε which defines the maximum allowable distance from face space (according to Equation 3.11).

5. For each new face image to be identified, calculate its pattern vector Ω, the distances ϵ_i to each known class, and the distance ε to face space. If the minimum distance $\epsilon_k < \theta_\epsilon$ and the distance $\varepsilon < \beta_\varepsilon$, classify the input face as the individual associated with class vector Ω_k. If the minimum distance $\epsilon_k > \theta_\epsilon$ but distance $\varepsilon < \beta_\varepsilon$, then the image may be classified as "unknown", and optionally used to begin a new face class.

6. If the new image is classified as a known individual, this image may be added to the original set of familiar face images, and the eigenfaces may be recalculated (steps 1 – 4). This gives the opportunity to modify the face space as the system encounters more instances of known faces.

Chapter 4

Experiments

Hofstadter's Law: *It always takes longer than you expect, even when you take into account Hofstadter's Law.*

Douglas Hofstadter, *Gödel, Escher, Bach*

4.1 Introduction

The performance of recognition algorithms are typically analyzed by one (or more) of three methods: (1) worst-case analysis, (2) probabilistic analysis, or (3) empirical testing. For most schemes based on well-defined features, such as corners of polyhedral objects, or those limited to well-defined object models — in CAD-based vision, for example — methods (1) and (2) are both possible and desirable. For complex object recognition tasks such as face recognition, however, analyzing performance is less straightforward. Because it is impossible to exhaustively catalog the range of objects expected, and because there is no clearly defined lowest-level feature set to work from (besides the actual pixel intensity values), analysis methods (1) and (2) are limited to particular data sets. These methods of performance analysis have little meaning on limited sets, so empirical testing becomes the dominant mode of performance analysis for complex object recognition. This chapter and the next focus on learning about the usefulness, limitations, and performance of the "eigenfaces" approach to face recognition from two approaches to empirical testing.

To initially assess the viability of this approach to face recognition described in Chapter 3 and particularly the objectives of accuracy and robustness, recognition experiments were performed on a set of stored face images, collected under a range of imaging conditions. Using this database I ran several experiments to evaluate the performance under known variations of lighting, scale, and head orientation. The results of these experiments are reported in this chapter.

4.2 Image database

The images from Figure 3-4(a) were taken from a database of over 2500 face images digitized under controlled conditions.[1] Sixteen subjects were digitized at all combinations of three head orientations, three head sizes or scales, and three lighting conditions. A six level gaussian pyramid was constructed for each image, resulting in image resolution from 512x512 pixels down to 16x16 pixels.[2] Figure 4-1 shows the images from one pyramid level for one individual. The subjects were allowed to move in between images, and were approximately but not exactly centered in the image. No attempt was made to precisely calibrate the imaging conditions beyond the gross distinctions in scale, lighting, and orientation.

To reduce the effect of the background on the calculation of the eigenfaces and the classification, the images were multiplied by a fixed gaussian window centered on the face, as shown earlier in Figure 3-10(b). The gaussian window emphasizes the center of the face and de-emphasizes the head outline, hair, and scene background.

4.3 Recognition experiments

In the first experiment the effects of varying lighting, size, and head orientation were investigated using the complete database of 2592 images of the sixteen individuals shown in Figure 3-4(a). Various groups of sixteen images were selected and used as the training set. Within each training set there was one image of each person, all taken under identical conditions of lighting, image size, and head orientation, all at the same scale. The top eight eigenfaces calculated from each training set were used in the classification process. Other face images from the database were then

[1] A subset of these images is available via ftp from "victoria.media.mit.edu" (net address 18.85.0.121), in the file pub/images/faceimages.tar.Z.

[2] So 16x3x3x3=432 images are unique, and the rest are filtered, subsampled versions of those. Altogether there are 432x6=2592 images.

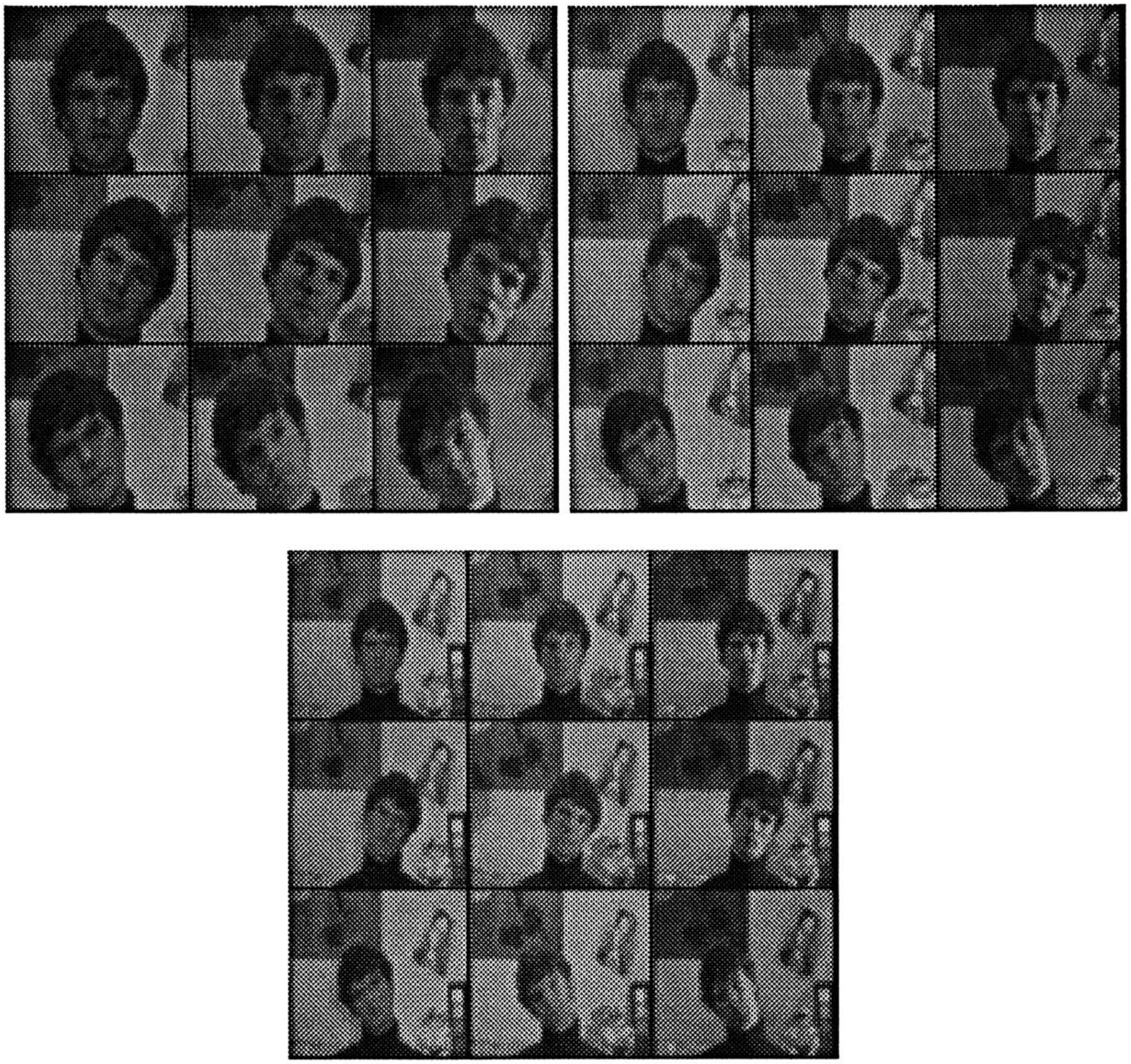


Figure 4-1: Variation of face images for one individual: three head sizes, three lighting conditions, and three head orientations.

classified as being one of these sixteen individuals — the one closest in face space using the euclidian distance metric — or else as "unknown". Statistics were collected measuring the mean recognition accuracy as the training conditions and the test conditions varied. The independent variables were difference in illumination, imaged head size, head orientation, and combinations of illumination, size, and orientation.

Figures 4-2 and 4-3 show results of these experiments. The graphs indicate the percentage of correct classifications for varying conditions of lighting, size, and head orientation, and combinations thereof, averaged over the number of experiments. The results are plotted as a function of the rejection rate, the percentage of faces rejected as unknown, which is controlled by the threshold parameter θ_ϵ (see Section 3.4). A rejection rate of zero is effected by an infinite threshold θ_ϵ. In this case where every face image is classified as known, the system achieved approximately 88% correct classification averaged over lighting variation, 65% correct averaged over orientation variation, and 55% correct averaged over size variation. Note that for the database size of sixteen, random chance should produce a 6% correct classification rate.

At low values of θ_ϵ (i.e. higher rejection rates), only images which project very closely to the known face classes will be recognized, so that there will be few errors but many of the images will be rejected as unknown. At high values of θ_ϵ most images will be classified, but there will be more errors. Adjusting θ_ϵ to achieve 98% accurate recognition boosted the unknown rates to 42% while varying lighting, 56% for orientation, and 59% for size. For varied lighting, a 93% recognition accuracy was reached with just a 14% unknown rate.

As can be seen from these graphs, changing lighting conditions causes relatively few errors, while performance drops dramatically with size change. This is not surprising, since under lighting changes alone the neighborhood pixel correlation remains high, but under size changes the correlation from one image to another is largely lost. It is clear that scale must be taken into consideration. The head size must be estimated, using either motion processing or a multiscale approach, as was discussed in Section 3.6.2, so that faces of a given size are compared with one another.

These experiments show an increase of performance accuracy as the threshold decreases. This can be tuned to achieve very accurate recognition as the threshold tends to zero, but at the cost of many face images being rejected as unknown. The tradeoff between rejection rate and recognition accuracy will be different for each of the various face recognition applications. However it is most desirable to have a way

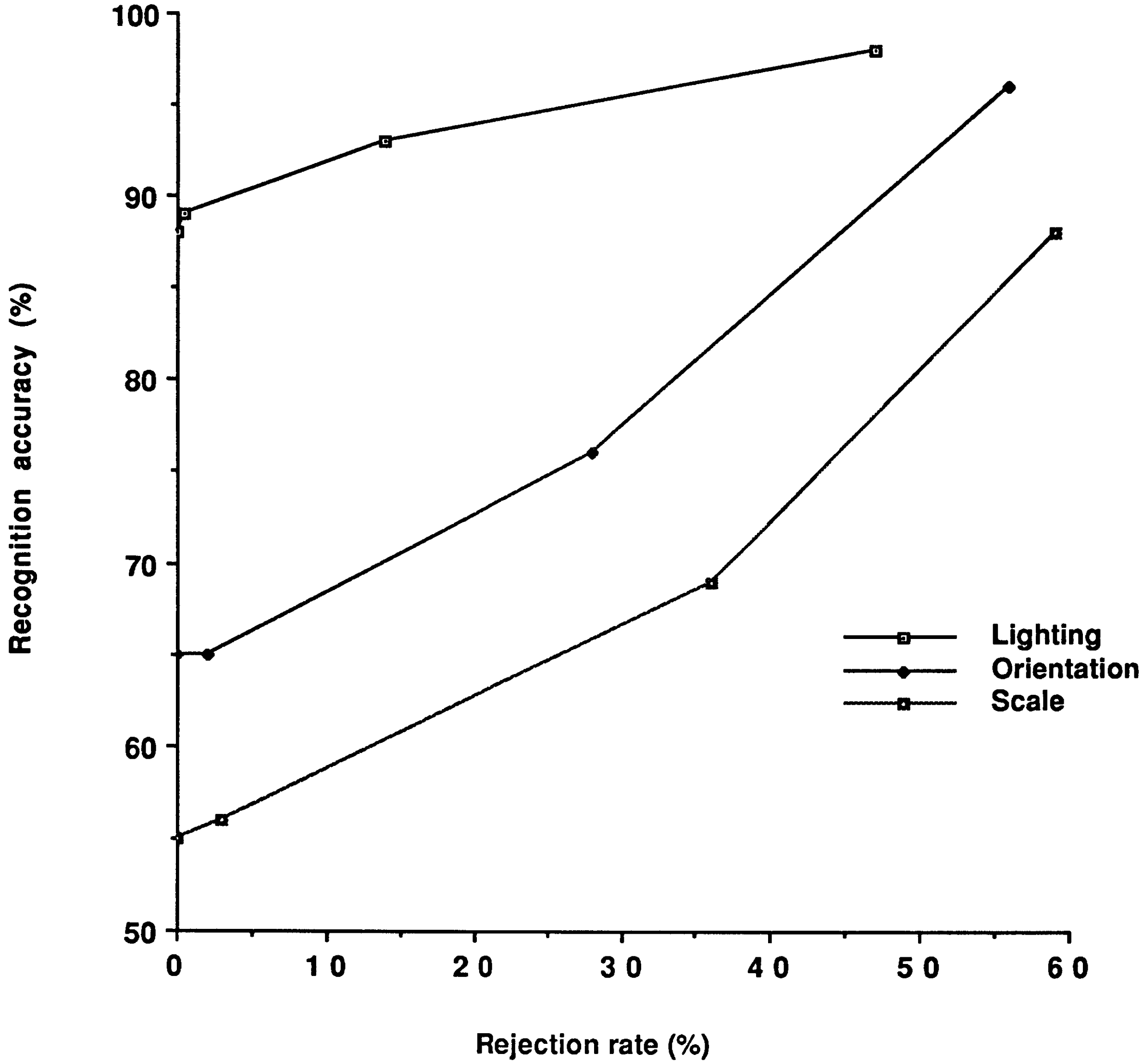


Figure 4-2: Results of experiments measuring recognition performance using eigenfaces, plotted as a function of the rejection rate. The averaged recognition performance as the lighting varied, as the head orientation varied, and as the head size (scale) varied.

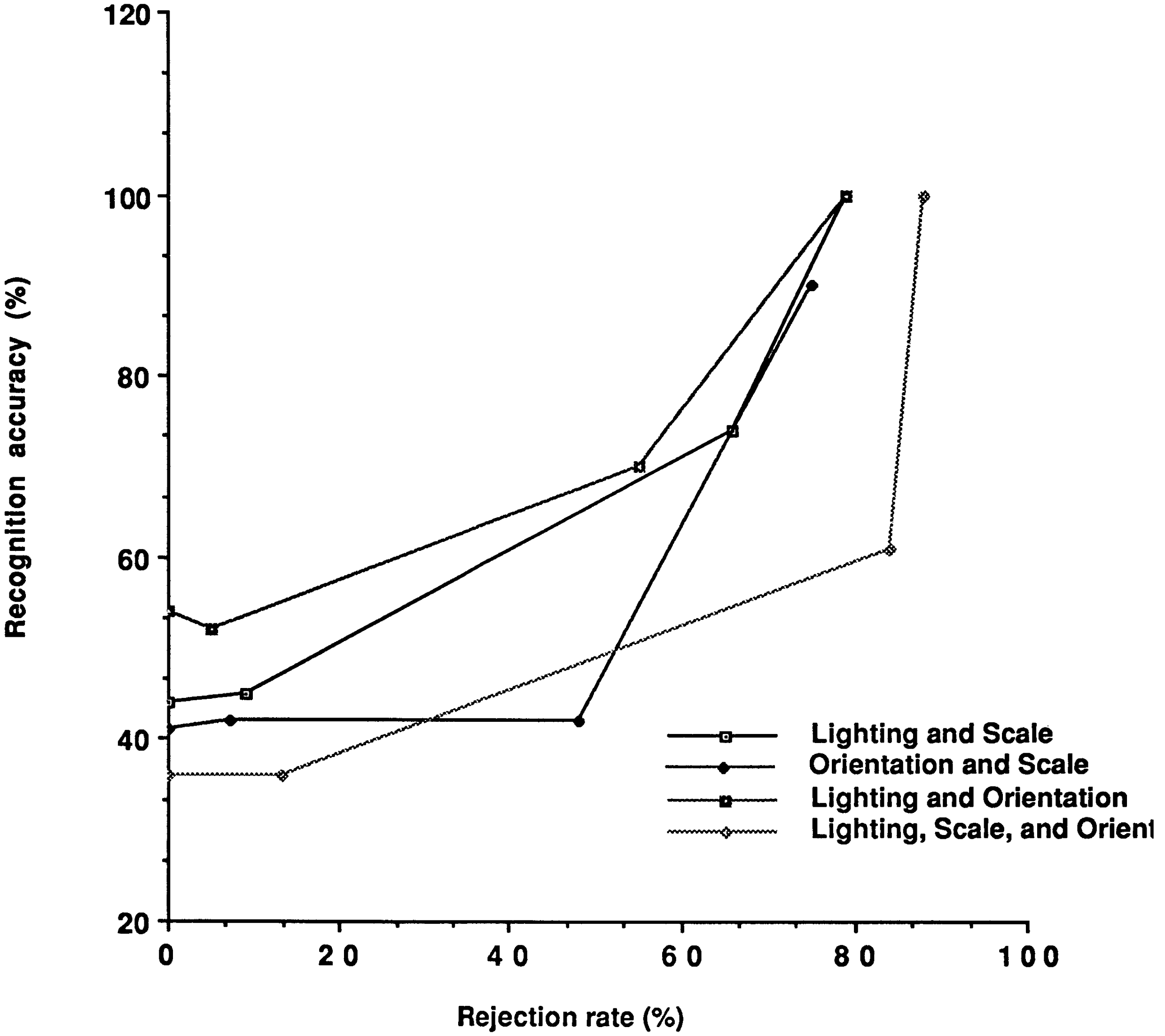


Figure 4-3: Results of experiments measuring recognition performance using eigenfaces, plotted as a function of the rejection rate. Recognition performance as combinations of the factors are varied.

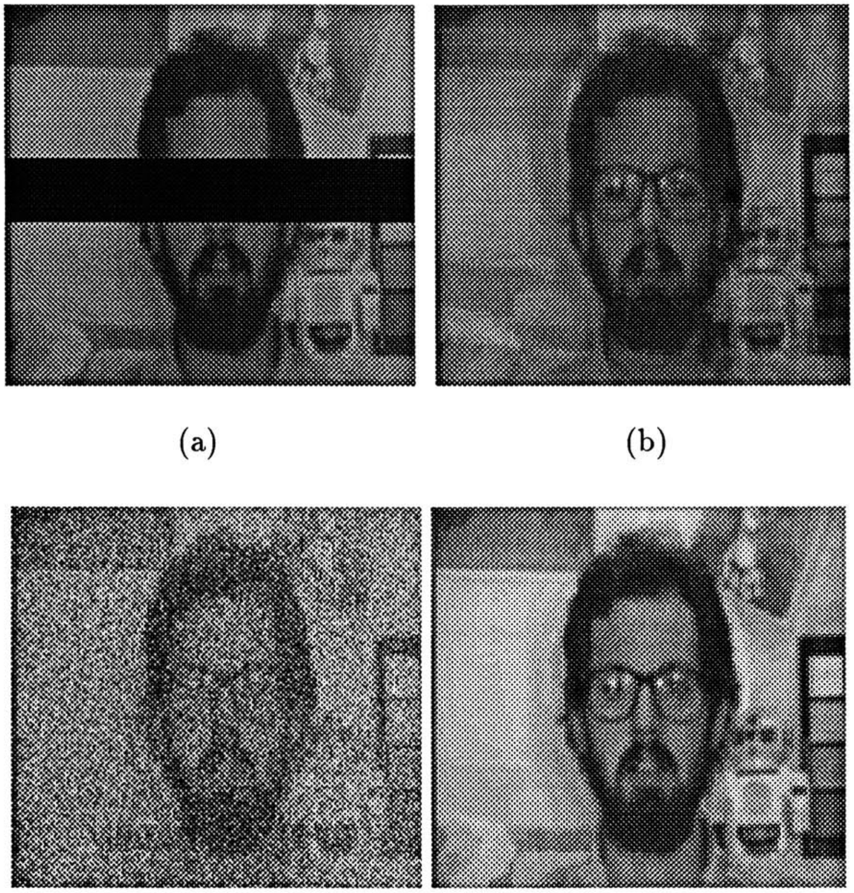
of setting the threshold high, so that few known face images are rejected as unknown, while at the same time detecting the incorrect classifications. That is, we would like to increase the discriminability (the d') of the recognition process.

The "distance from face space" metric (ε) introduced in Section 3.4 accomplishes this, allowing the rejection of false positives. Because the projection onto the eigenface vectors is a many-to-one mapping, there are a potentially unlimited number of images that can project onto the eigenfaces in the same manner, i.e., produce the same weights. Many of these will look nothing like a face, as shown in Figure 3-8(c). Although the experiments described in this section did not use this measure since they were all known to be faces, the recognition accuracy would certainly improve taking ε into account via the additional threshold β_ε (minimum distance from face space) — at the expense of an increased rejection rate.

4.4 Effects of artificial noise and the number of eigenfaces

Informal testing on face images with structured and unstructured (random) noise added in show that the system is reasonably robust to degraded input. (See Appendix A for a discussion of modeling noise in statistical pattern recognition tasks.) To be robust in the face of the types of image degradations depicted in Figure 3-1, a noisy image or partially occluded face should cause recognition performance to degrade gracefully, rather than abruptly. Because the eigenfaces essentially implement an autoassociative memory for the known faces (as described in [56]) local degradations are in a sense distributed throughout the image, as a local error will affect each eigenface projection a small amount. An example of this is shown in the occluded face image and face space projection of Figures 4-4 and 4-5. The face space projection effectively distributes the local error (the occlusion of the eyes) globally and recovers the missing information. The result of the occlusion is a larger distance from face space measure ε and an increased distance from the proper face class ϵ_k, but not an abrupt loss in the ability to recognize the face.

As mentioned in Chapter 3, the number of eigenfaces to use from a given training set is a choice that involves a tradeoff between recognition speed and accuracy. The heuristic rule which has worked well for the system is to choose the eigenfaces whose



(a)

(b)

(c)

(d)

Figure 4-4: (a) Partially occluded face image from the test set and (b) its projection onto face space. The occluded information is encoded in the eigenfaces. (c) Noisy face image and (d) its face space projection. (All images are recognized correctly.)

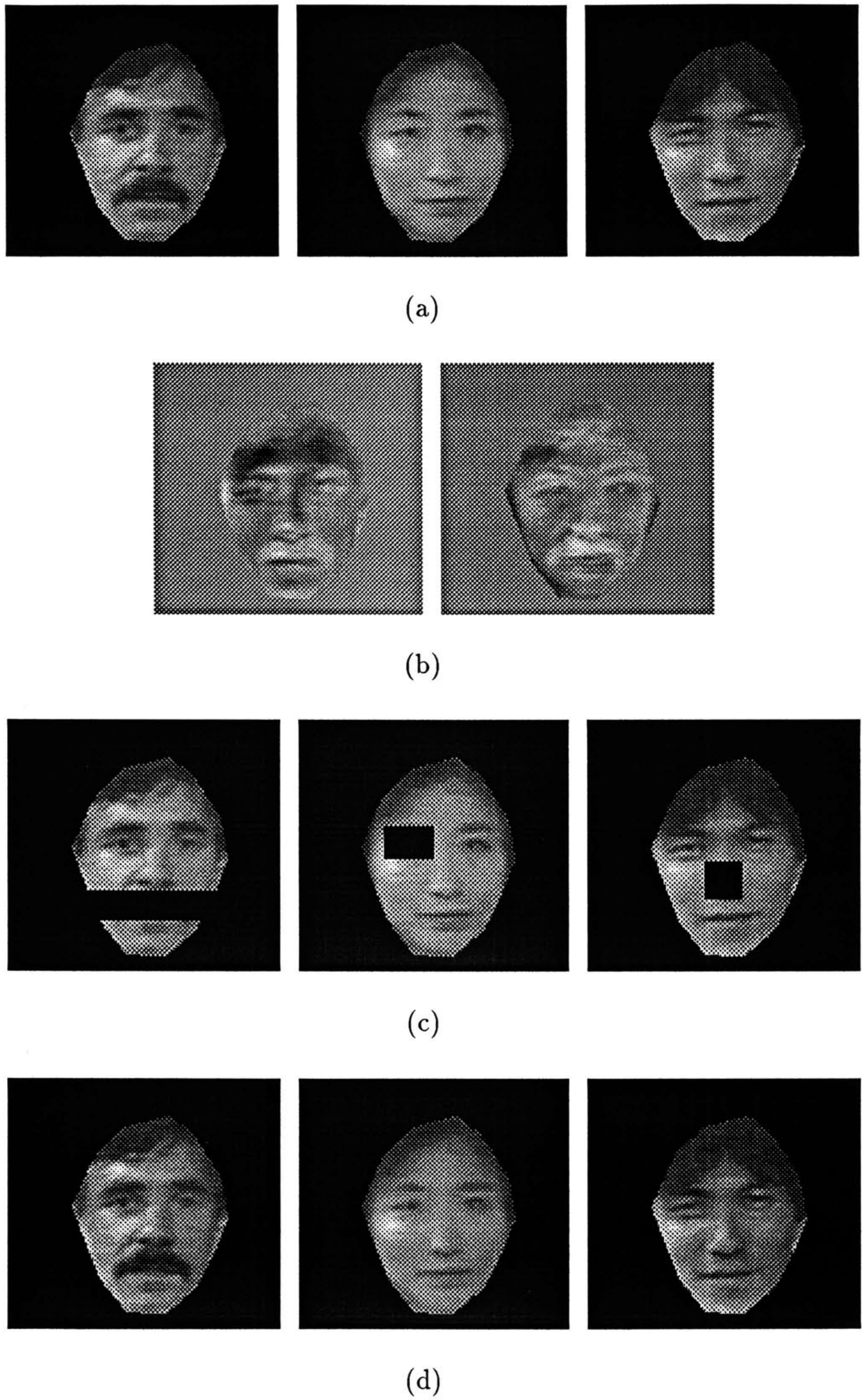


Figure 4-5: (a) A training set of three faces. (b) The corresponding eigenfaces. (c) Occluded images. (d) Their projections onto face space.

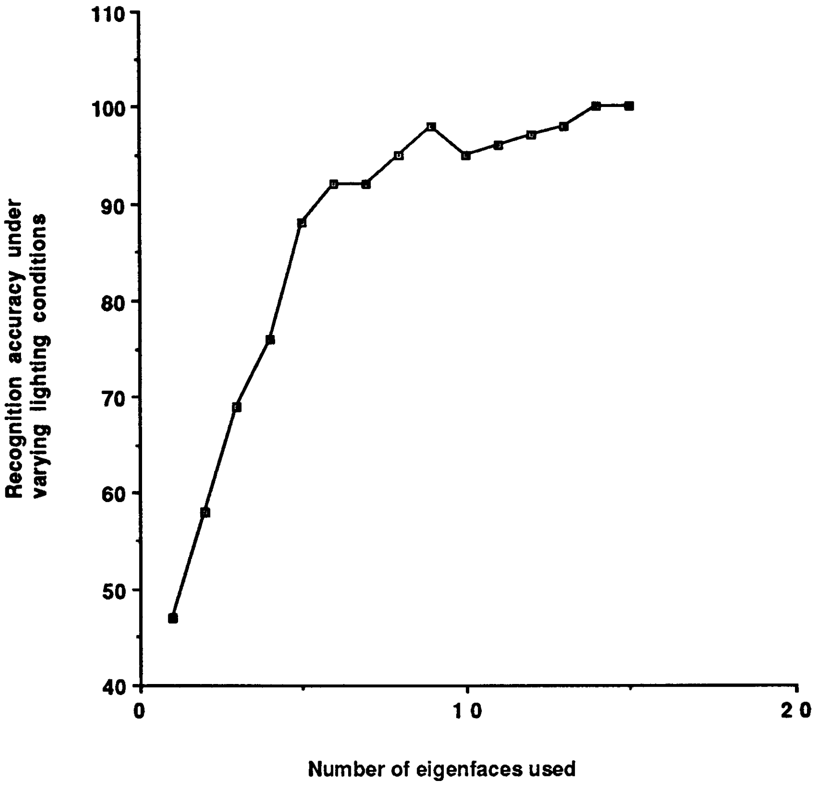


Figure 4-6: Recognition performance depends on the number of eigenfaces used. The graph shows the recognition accuracy over lighting changes for a training set of sixteen faces.

eigenvalues are within an order of magnitude of the largest. Figure 4-6 shows a graph of on test of recognition accuracy over a set of lighting changes as a function of the number of eigenfaces used. For this training set of sixteen faces, the performance drops below 90% at about five eigenfaces.

4.5 Lessons

Though somewhat impressive, the recognition results reported in this chapter should be taken as qualitative, not quantitative, performance measures of the approach. Many factors that were not controlled for precisely contributed to the results. Initially the background was not taken out for the experiments, so the whole image was being "recognized". This had both positive and negative effects, since although the background remained similar for the images of any one individual — which should make recognition easier — the background also affected the calculation of the eigenfaces which made recognition more difficult. The gaussian mask was chosen to reduce the background because of its simplicity and its deemphasis of the hair.

In addition, the subjects were allowed to move in between images, and some moved much more than others. Although the faces were approximately centered in the digitized images, the centering was done only by eye (mine, that is), so many of the images were off-center by a number of pixels. With the face location technique of Section 3.5, the faces would be much better localized, and recognition should be more accurate.

Another consideration is that because there was only one image grabbed under each condition, in every experiment one of the sets being tested was the training set itself. There would have been more room for error if for every training set there was another set of sixteen images digitized under the same conditions. However, informal tests indicate that there would have been little difference with training sets of this size.

The primary lessons from these experiments are:

1. The recognition approach is viable.
2. The system is accurate, comparable with published results of other systems.
3. Tradeoffs can be made between accuracy and rejection rate which will depend on the application.
4. A good estimation of scale is important for accurate recognition.
5. The background must be reduced or eliminated.
6. The system can handle reasonable amounts of occlusion and noise.

Chapter 5

Interactive-Time System

FACE RECOGNITION: No one has yet been able to build vision machines that approach our human ability to distinguish faces from other objects — or even to distinguish dogs from cats. This remains a problem for research.

Marvin Minsky, *The Society of Mind*

5.1 Introduction

To further test the performance of the approach to face recognition presented in this thesis — and to accomplish the goals of the work itself — it was important to implement the recognition in an interactive-time system. The main tasks of the face recognition system are to (1) determine that a face is present, (2) locate the face, and (3) identify the face. The recognition procedure based described in Chapter 3 uses the "face map" $\varepsilon(x,y)$ to perform the first two, and the "face space" classification to do the identification. However the face map from Section 3.5 is rather computationally expensive (especially when scale is unknown) and may currently be most practical in limited regions of the image, when the *approximate* location of the face is known, as well as in an initial bootstrap mode to initially find potential face locations. Therefore a faster, simpler technique for finding the location of faces is desirable.

Using the techniques described in Chapter 3 together with a simple motion detection and tracking algorithm, I built a system which quickly locates and recognizes faces in a reasonably unstructured, dynamic environment. Figure 5-1 shows a block diagram of the complete system. A fixed camera, monitoring part of a room, is con-

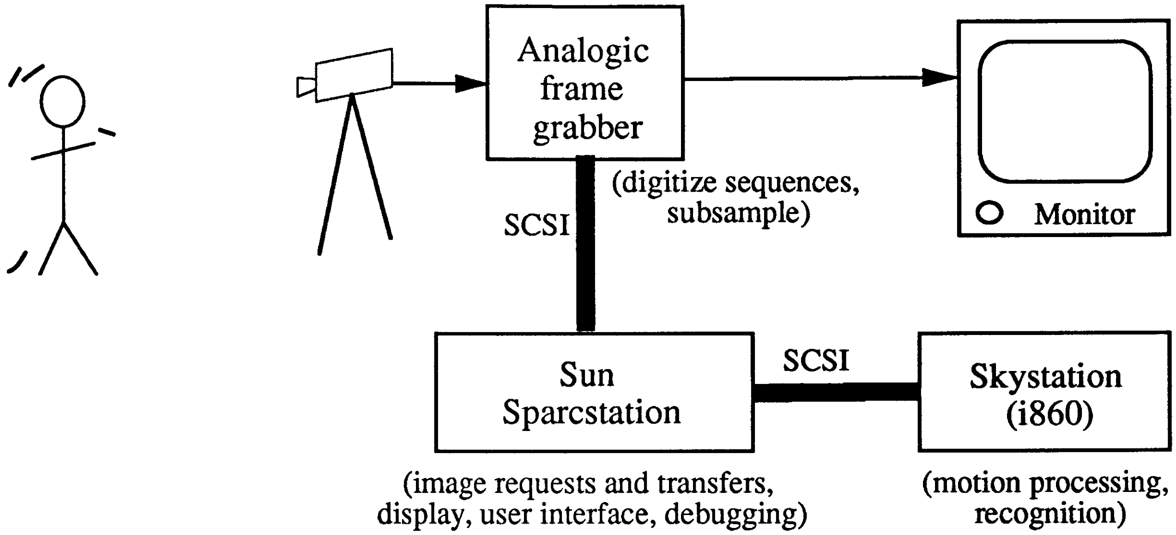


Figure 5-1: System diagram of the face recognition system.

nected to an Analogic frame grabber, which is attached to a Sun Sparcstation through the SCSI port. A Skystation, an i860-based application accelerator, is connected to the Sun and performs all the number crunching. The frame grabber digitizes and subsamples a sequence of video images at frame rate (30 frames/sec), and ships a subset of these 120x128 images to the Sun/Skystation for motion processing. If a head is located, a portion of the full frame (480x512) image is requested from the frame grabber to be used in the recognition.

The motion detection and analysis program looks for a moving object against a stationary background by tracking the motion and applying simple rules to determine if it is tracking a head.[1] When a head is found, a subimage, centered on the head, is sent to the Sun/Skystation. When the "face space map routine" is activated, the subimage is 256x256 pixels and the $\varepsilon(x,y)$ map is calculated to determine the exact location of the face. Otherwise, the face subimage determined from motion alone is shipped from the frame grabber. Using the distance from face space measure, the proposed face subimage is either rejected as not a face, recognized as one of a group

[1] Because of hardware limitations, the current system looks for motion only occasionally, whereas the first system, which used Datacube image processing hardware, tracked almost continuously. The current hardware was chosen for its simplicity, cost, and compactness, and is not best suited for the implementation.

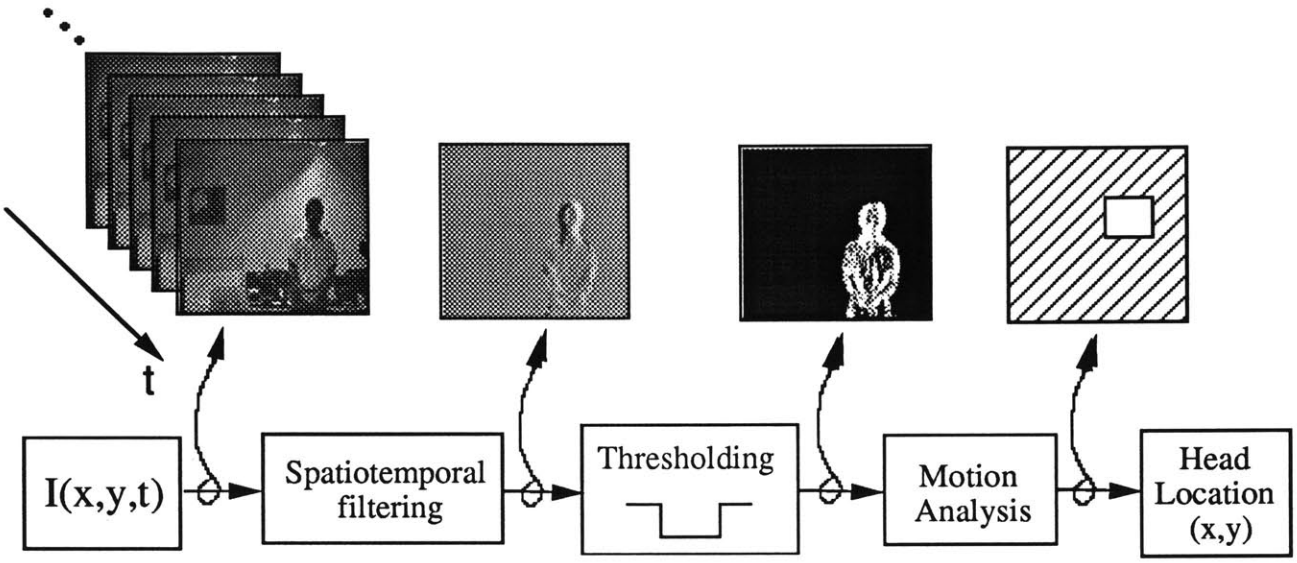


Figure 5-2: The head tracking and locating system.

of familiar faces, or determined to be an unknown face.

5.2 Motion detection and head tracking

Because people are constantly moving — even while sitting, we fidget and adjust our body position, nod our heads, look around, and such — motion can be a useful cue in estimating head position. In the case of a single person moving in a static environment, the motion detection and tracking algorithm depicted in Figure 5-2 will locate and track the position of the head. Simple spatio-temporal filtering (filtering, subsampling, and frame differencing) accentuates image locations which change with time, so a moving person "lights up" in the filtered image. If a significant portion of the motion map is above threshold, motion is detected and the presence of a person is postulated. (A similar motion detection algorithm which can deal with multiple objects and a moving camera has recently been demonstrated by Nelson [67].)

After thresholding the filtered image, the "motion blobs" of the binary motion map are analyzed to decide if the motion is caused by a person and if so to determine the head position. A few simple rules are applied, such as:

1. Small, isolated motion blobs are removed.

2. Too much motion in the image indicates that the threshold was set too low or there were many moving objects or perhaps camera motion. In this case, processing is aborted.

3. Head motion must be realistically slow and contiguous, since heads aren't expected to jump around the image erratically. This is implemented by filtering the path of head motion.

4. The head is assumed to be the upper motion blob, or the upper part of the large blob. (Headstands are not allowed!)

5. If no significant motion is detected, assume the head has not moved since it was last detected.

Figure 5-3 shows an image with the head located, along with a trace of the supposed path of the head in the preceding sequence of frames.

The motion map also allows for an estimate of scale. The size of the blob that is assumed to be the moving head determines the scale at which recognition is attempted. The motion map is expected to capture either the complete head or an outline including the sides of the head, both shown in 5-4. In both cases an estimate of head width — and therefore scale — is simple.

5.3 System considerations

Designing a practical system for face recognition within this framework requires assessing the tradeoffs between generality, required accuracy, and speed. If the face recognition task is restricted to a medium or small set of people (such as the members of a family or a small company), a small set of eigenfaces is adequate to represent the faces of interest. If the system is to reliably learn new faces or recognize many people, a larger basis set of eigenfaces will be required. The results of Sirovich and Kirby [92, 53] for coding of face images gives some evidence that even for large segments of the population, the number of eigenfaces needed is still relatively small. An intuitive explanation for this is that, as the number of faces in the training set grows, the rate of novel features or configurations seen (e.g. different kinds of noses,

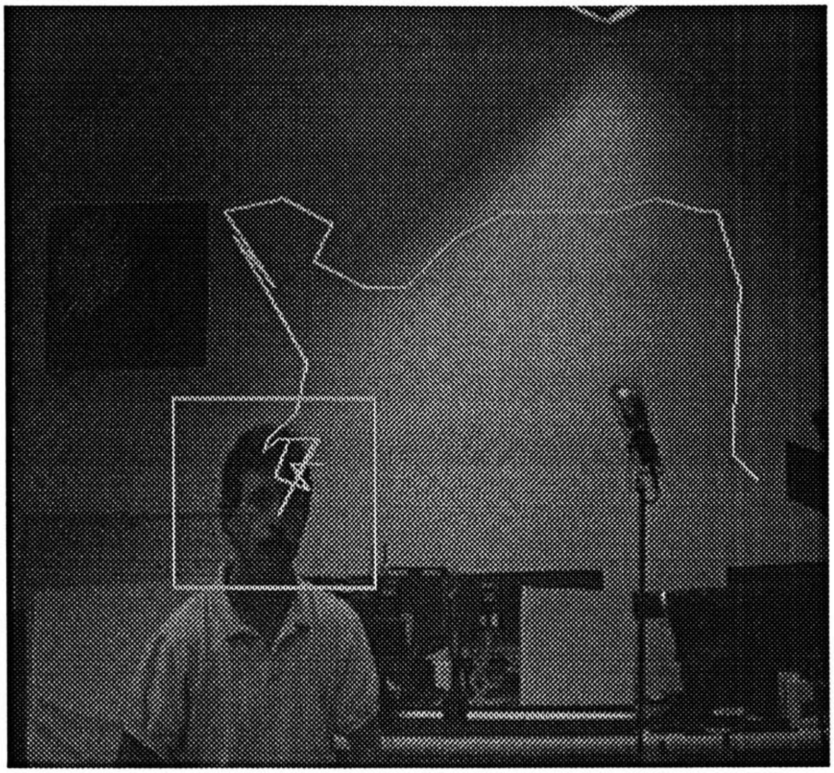


Figure 5-3: The head has been located — the image in the box is sent to the face recognition process. Also shown is the path of the head tracked over several previous frames. (The subject entered the scene from the right, walked over and sat down.)

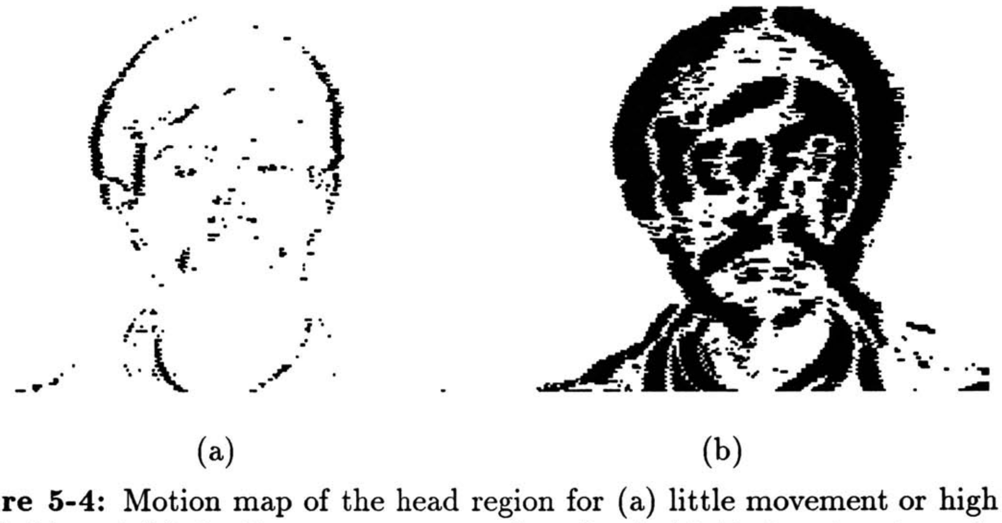


Figure 5-4: Motion map of the head region for (a) little movement or high threshold, and (b) significant movement or low threshold. Estimating the scale involves measuring the distance between right and left sides of the head.

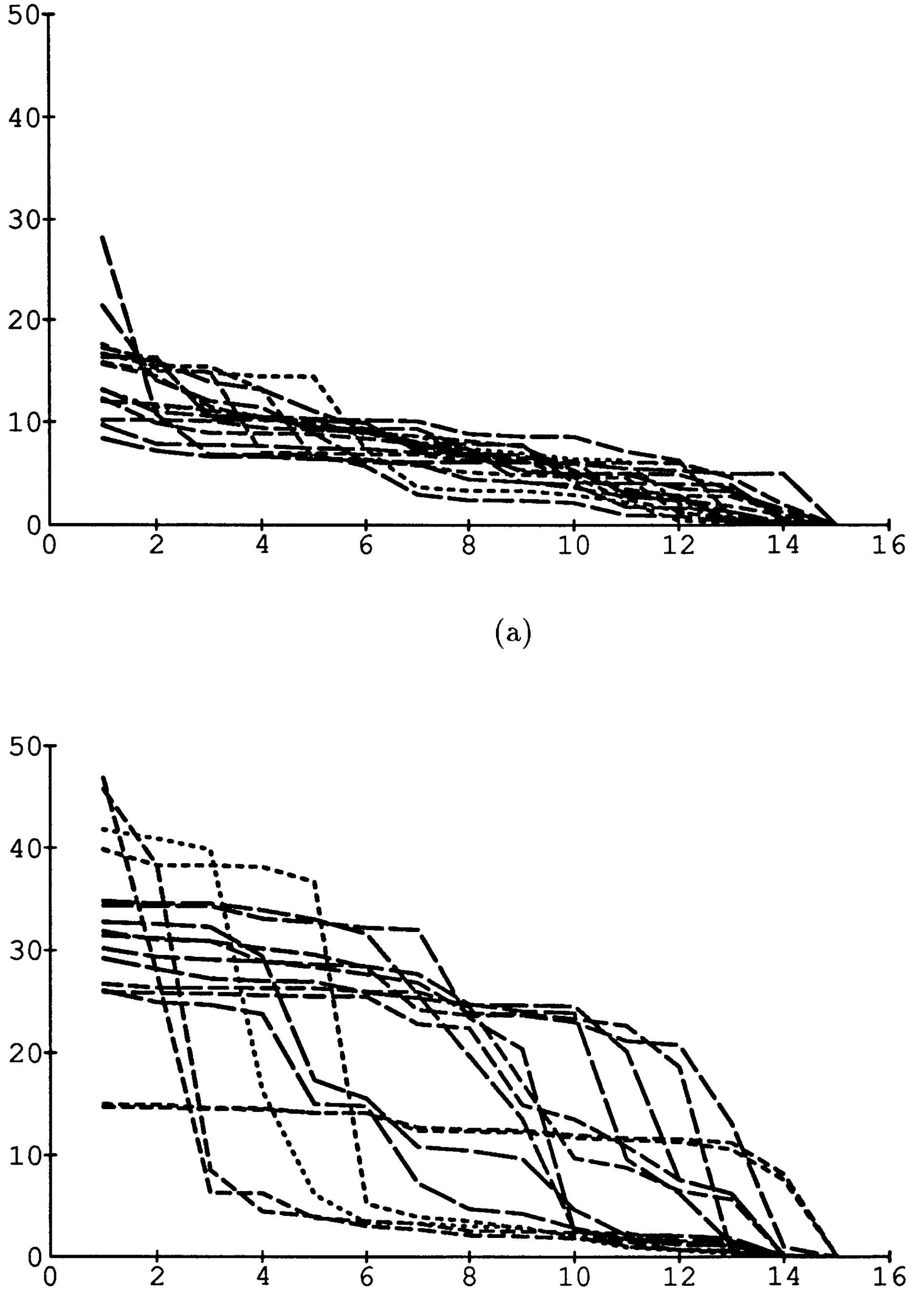
chin structure) decreases. So the first ten faces in the database may require six (for example) eigenfaces to capture their variation, while the last ten faces may require only one additional eigenface. Figure 5-5 shows graphs of reconstruction accuracy versus the number of eigenfaces used for (a) face images and (b) images of various unrelated objects. Clearly when the images share a common overall configuration, as do faces, there is a high degree of correlation among the images and fewer eigenfaces are needed to represent the class of images to a given error tolerance.

Even with a large number of eigenfaces, however, processing time may be speeded up by clever implementations of the calculations. For example, Equation 3.20 may be implemented so that after each image correlation with an eigenface $\mathbf{u}_i$ the temporary result is compared with the distance threshold β_ε, so that obviously non-face areas or images may be aborted long before all the eigenfaces are used.

As in the experiments of Chapter 4, the threshold θ_ϵ, which describes the maximum acceptable distance from the best face class, may be adjusted, as well as β_ε, the minimum acceptable distance from face space. These allow control over the accuracy rate and the false positive rate, respectively. A small θ_ϵ indicates that only very certain identifications should be made, those which fall quite close to a known face class, thus resulting in a higher rejection rate. A small β_ε means that only images very well accounted for by the eigenfaces will be considered for identification, resulting in few false positives but a higher sensitivity to noise or small changes in the face image.

The speed of the system depends on the options set, and is limited with the current hardware by the number of data and status transfers necessary over the SCSI port. Recognition occurs in this system at rates of up to a few times per second when the face map $\varepsilon(x, y)$ calculation is not performed and the face is found by motion analysis alone. Calculating $\varepsilon(x, y)$ currently slows it down to a rate of once every few seconds. Until motion is detected, or as long as the image is not perceived to be a face, there is no output. If motion is no longer detected, it is assumed that the individual is still and the face is expected to be in the same area as previously.

Using the system involves a short setup procedure, where the operator executes a few keyboard commands to digitize the people in the training set (or read the images from files), and uses the mouse to outline the head region for one image or else to locate the center of each face for multiplication by a gaussian window (e.g. see Figure 3-10). The calculation of eigenfaces is done offline as part of the training, and takes about fifteen seconds for a training set of sixteen people. After the training



(a)

(b)

Figure 5-5: Reconstruction accuracy for eigenfaces for (a) a set of sixteen face images, and (b) a set of sixteen images of unrelated objects. The graphs show the RMS reconstruction error (based on 8-bit intensity values) for each image in the training set as a function of the number of eigenfaces used.

set is collected, the system calculates the eigenfaces, defines the face classes (Ω_k), and begins the motion processing to look for potential face locations. When a face is recognized, the name of the identified individual is displayed on the video monitor.

Chapter 6

Biological Implications

> *Those who work in the field of artificial intelligence (AI) cannot design a machine that begins to rival the brain at carrying out such special tasks as processing the written word, driving a car along a road, or distinguishing faces.*
>
> David H. Hubel, *Eye, Brain, and Vision*

6.1 Biological motivations

High-level recognition tasks are typically modeled as requiring many stages of processing, e.g., the Marr paradigm [64] of progressing from images to surfaces to three-dimensional models to matched models. However the early development [1] and the rapidness of face recognition, along with the performance and selective nature of the neurological dysfunction prosopagnosia and the physiological studies discussed in Chapter 2, make it appear likely that there is also a recognition mechanism based on some fast, low-level, two-dimensional pattern recognition. Whether exclusively specific to faces or not, such a face recognition mechanism is plausible because of the nature of the visual stimulus (faces are typically seen in a limited range of views and orientations) and the social importance of face processing.

The approach and algorithms developed for face recognition in the previous chapters, then, are at least superficially relevant to biological vision. Without claim-

[1] A number of studies confirm that infants have a preference for face-like patterns. [41]

ing that biological systems store eigenfaces or process faces in the same way as the eigenface approach, we can note a number of qualitative similarities between our approach and both human performance and current understanding of the physiology. For example, the interrace effect — in which performance on face recognition tasks is demonstrably worse for faces of people who are of a different race than the subject — may be explained by the relative inadequacy of a face space constructed from experience with primarily one race or face type. In general, the system approach of motion processing to detect the presence of a face, calculating the face map to locate its precise location, and using eigenfaces to classify the face is similar to the typical human scheme of motion detection, foveation, and recognition.

Furthermore, as in human performance, relatively small changes cause the recognition system to degrade gracefully, so that partially occluded faces can be recognized. Gradual changes over time (e.g. due to aging) are easily handled by the occasional recalculation of the eigenfaces, so that the system is quite tolerant to even large changes as long as they occur over a long period of time. Similarly, human face recognition ability is invariant to gradual changes in appearance. If, however, a large change occurs quickly — e.g., addition of a disguise or shaving a beard — then the eigenfaces approach may be fooled, as are people in conditions of casual observation. The application of facial makeup, which should not affect feature-based approaches since makeup does not change the positions or relationships between features, will effect small changes in the eigenface representation — just as it causes subtle changes in perceiving a face.

If we consider the projection of an image onto a given eigenface analogous to the output of an "eigenface cell", the performance of that cell and of collections of such cells would be similar to many of the properties of face neurons in monkey cortex. For example, most face neurons respond differently to different faces, but respond to some degree independent of identity. Most are view-dependent. Most respond despite some occlusion, and respond more to the whole face rather than to any one part. These are also properties of eigenfaces.

There are a number of differences as well, which serve to highlight the fact that "eigenface cells" cannot account for the variety of documented face cells. Compare Table 6.1, which lists properties of "eigenface cells" with Table 2.1, listing properties of monkey face cells, to note particular similarities and differences.

Property	**Results**
color	gray-scale only, sensitive to luminance
orientation	not very sensitive to small orientation changes (approx. ±10°(rotation in the viewing plane)
position	dependent on centered face
size	very dependent on size of face
contrast	relatively invariant to small contrast changes (lighting)
identity	any given eigenface may respond the same or differently to different faces — depends on the current set
identity	is encoded not in individual eigenfaces but in the collection of their responses
expression	not very sensitive to small changes in expression
face view	view-dependent, but relatively insensitive to a small range about a given viewpoint
face view	discrete characteristic views
occlusion	respond degrades gracefully with occlusion
features	respond more to the whole face than to any one part
features	rather insensitive to the presence of a particular facial feature only
features	scrambling the configuration of features changes eliminates the response

Table 6.1: Properties of hypothetical "eigenface cells", which respond according to the projection of the image onto the eigenface

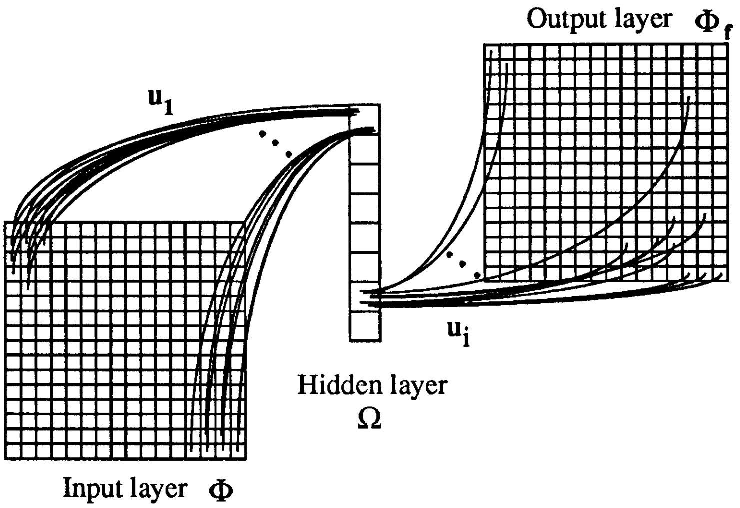


Figure 6-1: Three-layer linear network for eigenface calculation. The symmetric weights $\mathbf{u}_i$ are the eigenfaces, and the hidden units reveal the projection of the input image Φ onto the eigenfaces. The output Φ_f is the face space projection of the input image.

6.2 Neural networks

The hypothetical "eigenface cells", while not necessarily biologically plausible, may be learned by linear or non-linear units of a neural network trained using backpropagation [36]. Although I have presented the eigenfaces approach to face recognition as an information-processing model, it may be implemented using simple parallel computing elements, as in a connectionist system or artificial neural network. Figure 6-1 depicts a three-layer, fully-connected symmetric linear network which implements a significant part of the recognition system. The input layer receives the input (centered and normalized) face image, with one element per image pixel, or N elements. The weights from the input layer to the hidden layer correspond to the eigenfaces, so that the response of each hidden unit is the dot product of the input image and the corresponding eigenface: $\omega_i = \Phi^T \mathbf{u}_i$. The hidden unit responses, then, form the pattern vector $\Omega^T = [\ \omega_1\ \omega_2\ \dots\ \omega_L\ ]$. These units correspond the the "eigenface cells" discussed above.

If the output weights are symmetric with those of the input, the output layer produces the face space projection of the input image. This network implements an auto-associative memory, as described by Kohonen [55, 56], who offered a simple learning rule to modify the initially random connection weights. This network can recall noisy or partially occluded versions of the training set, identical to the behavior of the eigenfaces as shown in Figures 4-4 and 4-5.

The network by itself cannot perform recognition, but only produce output images from the input. The hidden layer of eigenface units must be fed into another network which can classify their outputs, and the output image — the eigenface reconstruction of the input — must also be fed to a network to determine the distance from face space, ε.

Adding two non-linear components we construct Figure 6-2, which produces the pattern class Ω, face space projection Φ_f, distance measure ε (between the image and its face space projection), and a classification vector. The classification vector is comprised of a unit for each known face defining the pattern space distances ϵ_i. The unit with the smallest value, if below the specified threshold θ_ϵ, reveals the identity of the input face image.

This network is used as a spotlight onto the scene — any patch which falls on its input image will be evaluated for "faceness" and identity. It therefore performs foveal face recognition. Together with a mechanism for drawing attention to possible face locations, perhaps via motion detection, the network could be implemented in hardware as a very fast face recognition system.

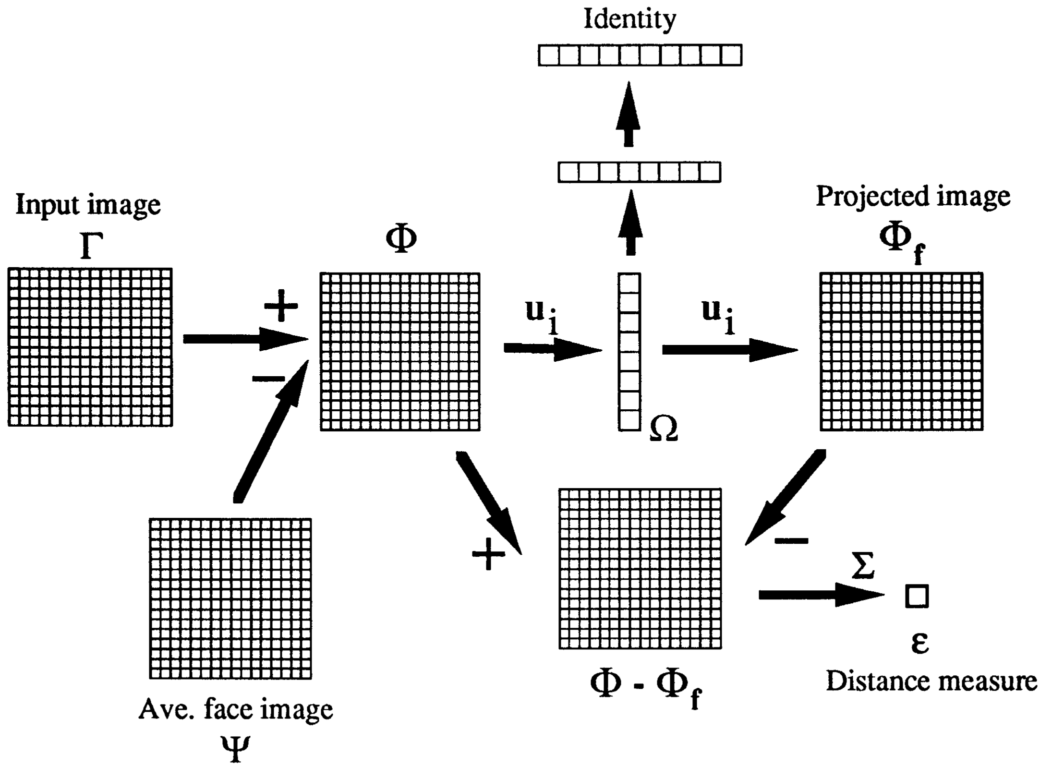


Figure 6-2: Collection of networks to implement computation of the pattern vector, projection into face space, distance from face space measure, and identification.

Chapter 7

Visual Behaviors for Looking at People

> *If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole, this is progress. ....Eventually this ... lands you with homunculi so stupid ... that they can be, as one says, "replaced by a machine." One discharges fancy homunculi from one's scheme by organizing armies of such idiots to do the work.*
>
> Daniel Dennett, *Brainstorms*

7.1 Introduction

Face recognition is of course not the only example of an interactive-time vision task devoted to human-computer interface. There are many useful tasks that may be approached through fast, reasonably simple visual "behaviors". An important factor in deciding what can be accomplished with this kind of system is to look at the tasks at hand and ask what pertinent information is available. Behaviors can then be constructed to take advantage of the nature of the task, to exploit a particular ecological niche.

In the particular area of human-computer interface, there are a number of visual behaviors that may be useful for machines to more intelligently and naturally interact with people. In addition to identity, people use many different visual cues in normal conversation to convey (or perhaps betray) information. In this chapter I will briefly describe a few of these which have been implemented to various degrees as visual behaviors. They are primarily for the purpose of demonstration, as they have not been extensively tested.

7.2 Feature detectors and direction of gaze

Knowledge of the presence or location of individual features, although probably not sufficient for the purposes of recognition, may be useful in other contexts or as components of a recognition strategy. The theory behind the eigenface analysis can be applied more specifically to individual features, creating eigenpictures for each feature of interest. As discussed in Appendix A, when looking for one class of object (e.g. eyes) in an image the analysis is similar to a matched filter when just one eigenpicture is used. Using multiple eigenpictures for the detection of a feature improves the performance. Figure 7.2 shows the top few eigenpictures for eyes, an original face image, and the "eye map" corresponding to the distance measure $\varepsilon(x, y)$. The dark spots indicate the eye positions. The eye detector is analogous to face recognition with just one known face class Ω.

Another useful visual behavior which has biological significance would be a "direction of gaze" detector. (Physiological studies [73] show many face cells in monkey cortex are highly sensitive to the direction of gaze of the observed face, some preferring eye contact, others preferring averted gaze.) Simpler than an eye tracker, which must output accurate direction information, a gaze detector should produce qualitative information concerning the viewing direction. The gaze detector knows three classes: left, right, and center. It is therefore a more general case of the above eye detector. Figure 7.2 shows an example of gaze detection. Both the eye detector and gaze detector can be implemented as interactive-time systems similar to that of Chapter 5.

7.3 Blink detection

Imagine a tiger crouching motionless along the jungle path, silently waiting for his prey to approach, preparing to pounce. Suddenly the prey notices in its periphery a

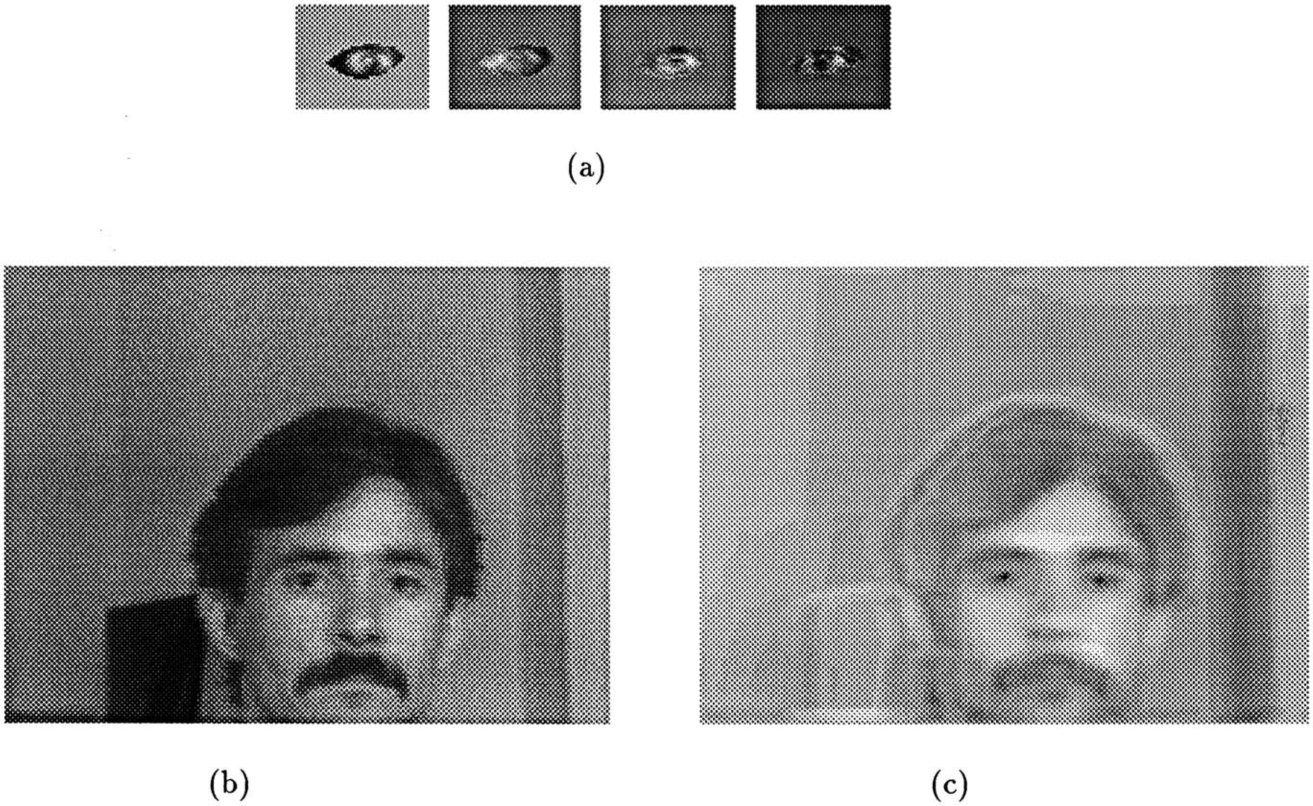


Figure 7-1: (a) Top four eigenpictures from a training set of eye images. (b) Face image. (c) Eye map $\varepsilon(x, y)$. The minima correspond to the eye positions.

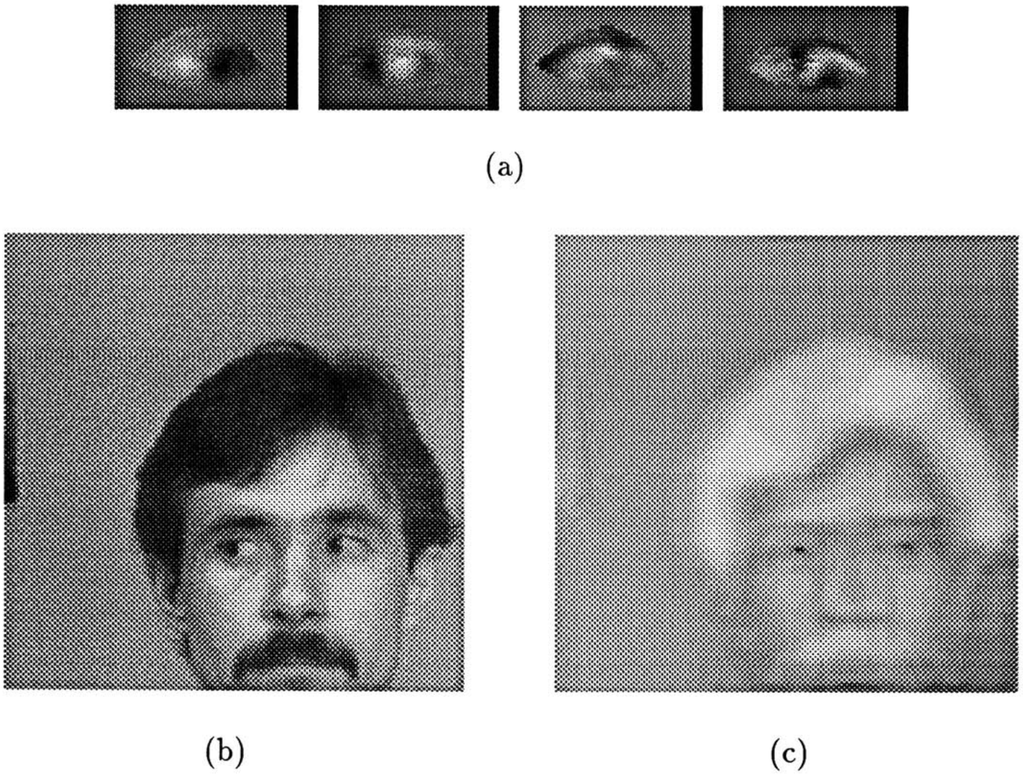


Figure 7-2: (a) Top four eigenpictures from a training set of eye images with direction of gaze to the left, right, and center. (b) Face image. (c) Eye map $\varepsilon(x, y)$. The minima correspond to the eye positions. At these positions, the subimages are correctly classified as gazing to the left of camera.

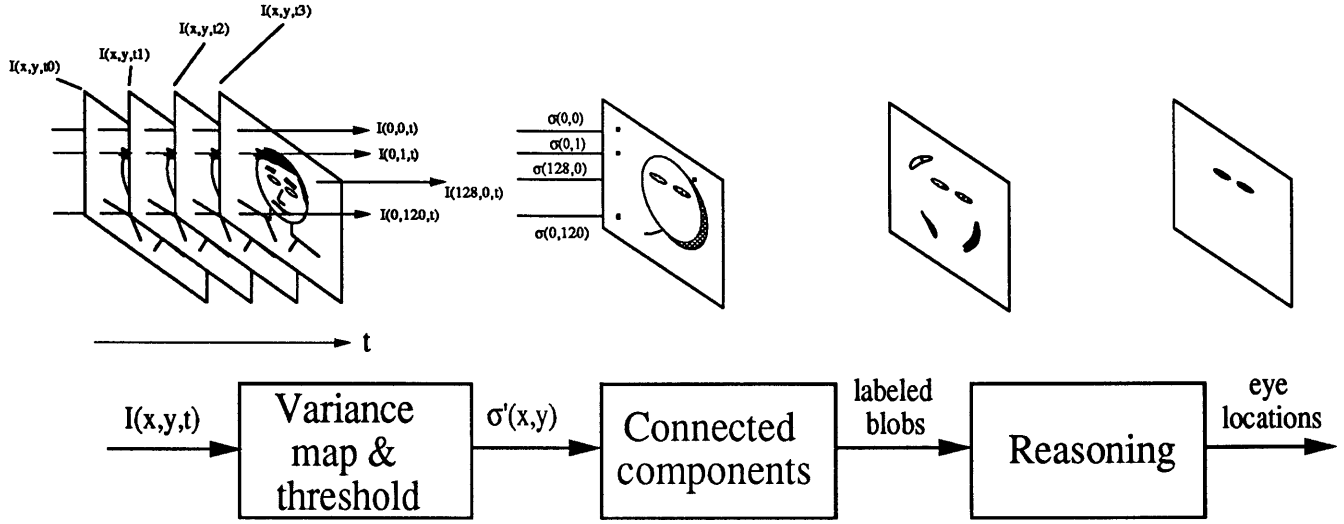


Figure 7-3: The blink detection algorithm.

small, brief, parallel movement and quickly darts off the path, escaping the imminent danger and frustrating the waiting predator. It may be stretching things to claim a biological need for mechanisms which detect eyes blinking, but because the information is available and well structured, it is relatively simple to build a visual behavior for blink detection.

Detecting eye blinks could be quite useful in the context of face recognition, helping to determine head location and orientation, scale, which image(s) from a sequence to use for identification (i.e. not to use an image with the eyes closed or closing), and possibly higher-level aspects of face processing such as identifying moods (e.g. people blink more often when they are nervous or embarrassed). Since blinking is performed consistently by (awake) people, detecting blinks may be a reliable and important tool for human-computer interface.

I have developed a simple technique for detecting eye blinks which has been tested on short motion sequences. The algorithm depends on intensity variations over a limited spatial and temporal region. The idea is depicted in Figure 7-3. Intensity variance over time is calculated on a pixel-by-pixel basis in a sequence of images, producing a "variance image" $\sigma(x,y)$. Pixels which have significant temporal variation, due to any of a number of factors — e.g. lighting changes, image noise, object motion, or moving shadows — will have a relatively large variance.

After each new image in the sequence, the variance image is checked for the number of pixels above a threshold θ_σ. If there are a sufficient number of "motion pixels", a connected components routine is applied to the thresholded variance image to find connected regions, and simple statistics are calculated for each region: the area in pixels, the centroid, and the minimum and maximum extent of the region. The regions are then analyzed by a simple rule-based routine to check for the presence of eye-blink pairs — pairs of regions which satisfy the following criteria:

1. Regions which are too large or small in area are discounted.

2. Regions which are far from circular are discounted.

3. Eye pairs must be approximately horizontal.

4. They must be within a certain range of horizontal spacing.

If there is general head motion, the routine will fail to find an adequate region pair and start over again, clearing the variance image.

Figure 7-4(a) shows selected images from a one second (30 frames) sequence of a person blinking, with very little other head movement. Figure 7-4(b) shows the variance image for a subset of the frames. Figure 7-4(c) shows the connected components of the thresholded variance image, and 7-4(d) the only candidate eye-blink pair. Obviously in this example the selected regions correspond to the eyes in the original image sequence.

This technique is quite simple and can be fooled by any small, parallel motion. Blinking during a movement of the head will not be detected, although compensating for this is reasonably straightforward using motion pyramids (see Nelson [67] and Burt [17] for fast techniques to detect small movements in the context of larger motion). However in brief periods of time when the head is still and there the eyes blink, the algorithm will "light up" the eyes as in Figure 7-4(e). Because in a great amount of communication and human-computer interaction people tend to limit their head motion, and because blinks are typically of a fixed duration in time, this simple approach is an effective visual behavior, well suited to its ecological niche.

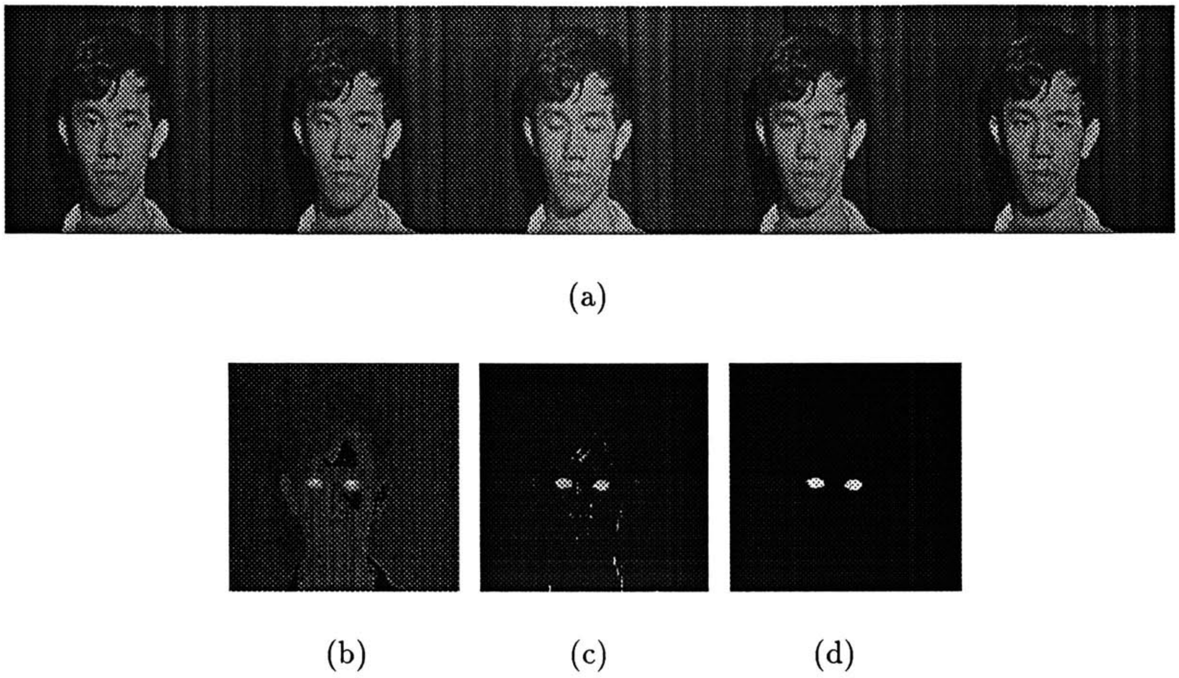


Figure 7-4: (a) Selected frames from a sequence of 30 images. (b) Variance image. (c) Connected components of the image from (b) after thresholding. (d) The single candidate eye-blink pair after the selection step.

7.4 Expression analysis

The eigenface approach may be useful for not only identification and detection, but for analyzing expression as well, particularly for a given individual. A training set was captured consisting of images of one person with a range of expressions. Classes were defined for the specific expressions of smiling and frowning. The system reliably distinguished between the two and exhibited the typical accuracy rate vs. rejection rate tradeoffs as in the case of face recognition. Figure 7-5 shows an example of the behavior.

It is evident from viewing the eigenfaces in this example that the moustache and eyebrows were significant. For some people, it may be that the teeth are more important indicators of expression. The eigenfaces let the system decide what to encode. Working in concert with the recognition system, the expression analysis can pull out the appropriate expression eigenfaces for the previously recognized individual.

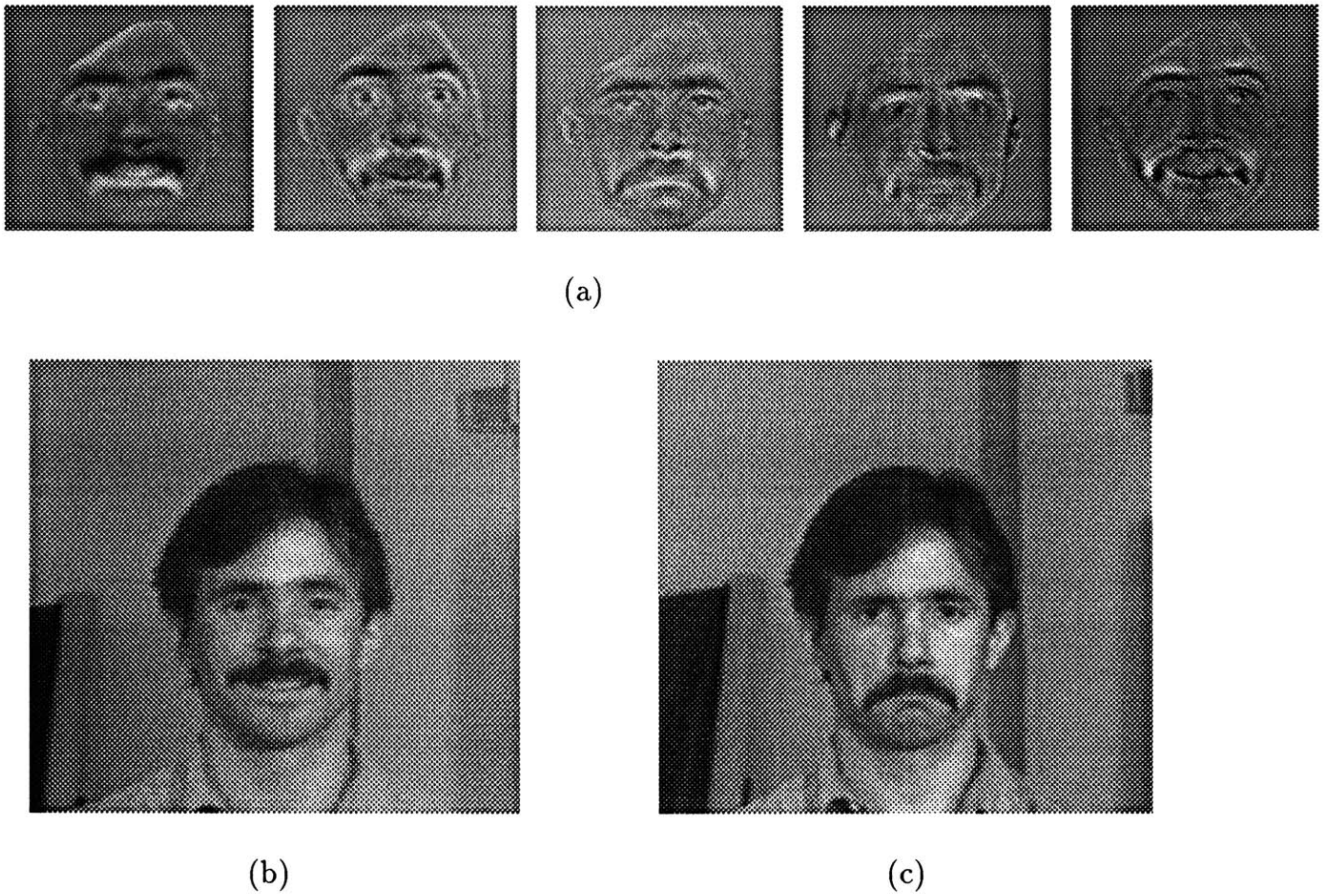


Figure 7-5: Detecting expression. (a) Eigenpictures from various expressions. (b) New face image classified as smiling. (c) New face image classified as frowning.

Chapter 8

Summary and Discussion

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get somewhere,*" Alice added as an explanation.*
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Lewis Carroll, *Alice's Adventures in Wonderland*

8.1 Interactive-time vision and visual behaviors

The previous chapters have described visual behaviors intended for applications of "looking at people": face recognition, feature detectors, blink detection, and expression analysis. Visual behaviors are interactive-time vision systems which implement special-purpose skills in a vertically layered fashion (as depicted in Figure 1-1). Face recognition is treated in depth, and is a particularly interesting visual behavior at this time because of the commercial interest[1], the feasibility in terms of speed and price of available hardware (both computers and optics), and the increased interest in biological face recognition and therefore computational models of recognition.

Interactive-time vision is a superset of active (or animate) vision, which is mainly directed towards behaviors relating perception to action by a robot. Along with the

[1] Many industries are very interested in face recognition systems for such applications as security systems, teleconferencing, picture processing, and people meters for TV ratings.

domain of active vision, interactive-time vision includes tasks in which a human completes the perception-action loop, and those which may only affect a robot's behavior occasionally rather than continuously. Thus the distinction between interactive-time and "real-time" is not just in the processing time, but in the manner in which the output of the system is used. To further clarify the distinction, a real-time stereo system which produces a depth map many times a second is an example of neither interactive-time vision nor active vision, since there is no goal achieved which would directly enable action. If that map is output to a display system which highlights all objects outside a safety area twice per second, the combination is an interactive-time system. If the highlighted objects are separately located to enable a machine to move them at a significant rate, the complete system is an example of active vision.

An important aspect of a visual behavior is the ability to report some measure of confidence along with its answer. Because a special-purpose system will not necessarily be operating in the ecological niche it is best suited for at all times, its output should often be suspect. In fact, such systems are designed to be wrong most of the time, and right only in special cases. Some indication of the validity of its answer — or the lack of an answer when the situation is not appropriate — is vital for these behaviors to be useful. In the case of face recognition, for example, the "distance from face space" measure ε gives an indication of whether a face is present at all, and the "distance from the nearest face class" measure ϵ_k indicates the identity confidence. Rather than merely requiring these values to be under some set thresholds, the values themselves should be output along with location and identity of the face(s). This allows for other behaviors to either supersede or defer to each other, depending on the relative confidences.

8.2 Face recognition via eigenfaces

The approach to face recognition described in this thesis meets the objectives stated in Chapter 3 of speed, accuracy, robustness, a limited ability to learn, and the ability to both locate and identify faces. From both experiments with a stored database of face images and experience with a working system, empirical results are promising and the system can recognize faces in a moderately limited domain. The following aspects of the system have not been fully implemented:

- Scale estimation — The thresholded motion map usually produces a clear silhouette of the head or the outline of the head. From this map the head width, and therefore the approximate scale of the face image, may be estimated. However because the motion is detected in a low resolution (120x128 pixels) subsampled image sequence in the current system, the head width measurement is likely to be up to a few pixels off. Since this measurement is used to rescale the higher resolution head subimage, any error in estimation at low resolution may be magnified. Motion analysis at high resolution is not feasible with the current hardware.

- Non-euclidian metric — The distribution in face space of a single face class is not necessarily uniform gaussian, which is assumed by the euclidian metric used for identification. Characterizing the distribution and using a non-euclidian classification scheme should improve the identification performance.

- Characteristic views — The system currently treats multiple, characteristic views of a single person as separate people, combining the views into one set of eigenfaces. There should instead be multiple sets of eigenfaces, one per viewpoint. Memory limitations of the Skystation application accelerator currently prohibit multiple eigenface sets.

Another aspect of the approach which has not been fully explored is its scalability, the performance as the face database increases in size. With the current system I have tested only databases up to twenty faces; however for various applications, databases of hundreds or even thousands of faces may be desirable. The main questions are:

1. How many eigenfaces are needed as the database size increases?

2. How does database size affect the reliable detection of faces?

3. How does database size affect the accuracy of face identification?

The first question is important since it affects the processing speed and memory requirements. As mentioned in Section 5.3, the image coding results of Sirovich and Kirby imply that as the database size grows, the number of eigenfaces needed to represent to ensemble of faces grows at a much smaller rate.

Of course the first question is related to the others, since the number of eigenfaces will affect recognition performance. The fewer eigenfaces used, the more the recognition performance deteriorates. Too few eigenfaces results in an inability to recognize a wide variety of faces as faces, since the distance from face space ε may be large even for those in the database. And the fewer eigenfaces there are, the fewer face space dimensions there will be, reducing the discriminability of identification.

The intuition gleaned from using the system indicates that the limiting factor will be the ability to discriminate face classes as the database grows very large. This can be viewed optimistically, however, as a "feature" rather than a "bug", in the following sense. For large databases such as an FBI collection of thousands of mug shots, the most useful application for a face recognition may be to use a photograph or artist's rendering to limit the number of likely suspects to manageable number, e.g. a page of mug shots. Because identification using the eigenfaces gives a distance measure to every face in the database, these can be ranked and the best N matches can be displayed for further inspection.

An alternative approach for very large databases is to do an initial screening as described above, and then calculate a new, temporary set of eigenfaces from the best N matches. The discrimination ability in this smaller set should be improved, and accurate identification should be much improved.

8.3 Extensions and other applications of eigenfaces

The approach to recognition using eigenfaces has no inherent knowledge about faces beyond masking out the background to only use the face region of the images. The technique should not be limited to faces; it should also be applicable to categories of objects which share the same overall configuration (as do faces) and are typically seen in a limited range of views. It has been suggested to use eigenfaces[2] to recognize or classify frontal views of cars or tanks, trees and bushes, and cursive writing, to name a few.

Additionally, the concept can be used more specifically within the context of face

[2] Or in these cases, "eigencars", "eigentrees", etc.

processing. Shackleton and Welsh [89] have begun to apply the eigenface analysis to individual facial features such as eyes, after first using deformable templates to find and normalize the features. This is effectively combining the work of Yuille *et al.* [108] with the work reported in this thesis. The approach is promising since it merges the holistic eigenface approach with the feature-based template approach to face recognition. Craw [26] also reports on face recognition work using face space, and propose to model the time-dependence of learning the faces using probabilistic techniques for calculating eigenvectors.

Choi *et al.* [23] report using the eigenface technique for both 3-D shape (by appropriately adjusting a wireframe model) and for recognition. Current face recognition research by Akamatsu [2] is strongly motivated as well by previous reports of this eigenface approach [99, 100, 101]. Other groups are also beginning to investigate the use of eigenface analysis to determine other categorical judgements besides identity, such as the subject's gender and facial expressions.

8.4 Suggestions for future research

There are many possible research directions and challenging problems involved in further improving the approach to face recognition introduced in this thesis. One such area is improving the invariance of the system to changes in lighting. The symmetry of the face gives the opportunity for adaptive filtering which will take away some of the effects of an uneven illumination source. Facial symmetry also gives a straightforward clue as to the head orientation, and I have briefly experimented with simple orientation operators to give an estimate of orientation (deviation from vertical).

Distinctions other than identity may be made based on eigenfaces, such as sex, race, age, and expression. As is the case in recent work of Cottrell and Metcalfe [24], it may be that a certain subset of the eigenfaces are most useful for any one of these evaluations. It is unlikely however that for large databases the eigenfaces which are best for discriminating identity will also be most useful for discriminating expression, since the former task seeks to map all the expressions of a given person into a single class. From the limited experiments of this thesis it seems that expression may best be analyzed by a separate set of "eigenexpressions" for each individual.

Computer graphics techniques may be used to artificially age the database and thus predict face classes or construct eigenfaces which are relevant to individuals who were seen years before. Similarly, patterns of facial hair may be predicted and rendered, merging imaging and graphics in the database so that most likely appearances of an individual are accounted for directly. For most of these tasks, the feature locations will need to be known precisely.

The examples of Figures 4-4 and 4-5 indicate that recognition may fare well when part of the face is occluded, or when there is significant image noise. These cases may be handled even better, improving the recognition accuracy, by using a two-step procedure: (1) project into face space, comparing the face image with the projection, and (2) throw away those pixels which are very dissimilar to the projection and classify the rest of the image. Scaled correctly, this should provide a more accurate identification, although research to determine its limitations is necessary.

Further work should be done on the application to very large databases, looking into how the problem scales with size and at what size individual identification becomes unreliable. This is important for at least two very significant applications of this work, locating faces in general scenes (e.g. for querying an image database) and reducing a victim's search in criminal identification.

In addition, the motion processing which precedes recognition should be extended to more general motion cases such as multiple moving objects and camera motion. Sophisticated motion processing can be vitally interconnected with the recognition scheme. For example, accurate head location reduces overall computation by reducing the image area of the "distance from face space" calculation.

Chapter 9

Conclusion

Of making many books there is no end, and much study wearies the body.
Now all has been heard; here is the conclusion of the matter.

Ecclesiates 12:12,13 (NIV)

The early attempts at making computers recognize faces were limited by the use of impoverished face models and feature descriptions (e.g. locating features from an edge image and matching simple distances and ratios), assuming that a face is no more than the sum of its parts, the individual features. Recent attempts using parameterized feature models and multiscale matching look more promising, but still face severe problems before they are generally applicable. Current connectionist approaches tend to hide much of the pertinent information in the weights which makes it difficult to modify and evaluate parts of the approach.

The eigenface approach to face recognition was motivated by information theory, leading to the idea of basing face recognition on a small set of image features that best approximate the set of known face images, without requiring that they correspond to our intuitive notions of facial parts and features. Although not intended as a solution to the general object recognition problem, the eigenface approach does provide a practical solution that is well fitted to the task of face recognition. It is fast, relatively simple, and has been shown to work well in a somewhat constrained environment.

Such a system implements a so-called "visual behavior", analogous to Brooks' vertically-layered robot behaviors or Minsky's cognitive agents. A visual behavior is an interactive-time system which solves a real recognition task in a particular ecological niche. Other simple examples of visual behaviors have been presented in the thesis, e.g. blink detection, feature detectors, and smile detectors.

It is important to note that many applications of face recognition do not require perfect identification, although most require a low false positive rate. In applications such as security systems or human-computer interaction, for example, the system will normally be able to "view" the subject for a few seconds or minutes, and thus will have a number of chances to recognize the person. For the task of searching a large database of faces it may be preferable — or at least more practical — to find a small set of likely matches to present to the user, rather than choosing just one face as the correct match. Our experiments show that the eigenface technique can be made to perform at very high accuracy, although with a substantial "unknown" rejection rate, and thus is potentially well suited to these applications.

The main technical contribution of this thesis is the development and implementation of a new approach to face recognition: using eigenfaces as a substrate for recognition, combining with simple motion processing to locate potential known or unknown faces, providing a capability to identify individuals or a group of likely candidates and to learn to recognize new faces in an unsupervised manner. The interactive-time nature of the system enables rapid experimentation and allows for a variety of useful tradeoffs such as recognition accuracy versus rejection rate. The experiments with a large database of face images under varying imaging conditions, as well as the ongoing experience with a working interactive-time system, lend significant experimental support to the approach.

Appendix A

Matched Filtering and Eigenfaces

Correlation techniques have been used for object recognition since the early years of computer vision research. These techniques work well in locating features ranging from simple edges and corners to complex shaded patterns [91], but typically only if the imaging conditions are well controlled and the features are well described by two-dimensional image patches. This appendix briefly discusses correlation as used in recognition, and its relationship to the eigenface-based recognition.

A.1 Correlation and matched filters

The task of object recognition is fundamentally one of comparison — comparing the available image data with some stored representation, however complex or simple, of known objects. One of the most basic and well-understood methods of object detection in image processing is template matching, where image intensities are compared directly. An image template $k(i,j)$ may be compared in a least mean squared sense by the distance measure

$$d^2(x,y) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I(x+i,y+j) - k(i,j))^2 \tag{A.1}$$

If a portion of the image $I(x,y)$ exactly matches the template the distance measure will be zero at the location of the match; otherwise it will be greater than zero, $d(x,y) \geq 0$. A match occurs when $d(x,y)$ is below some predetermined threshold, or alternatively a match is defined at the location of the minimum $d(x,y)$.

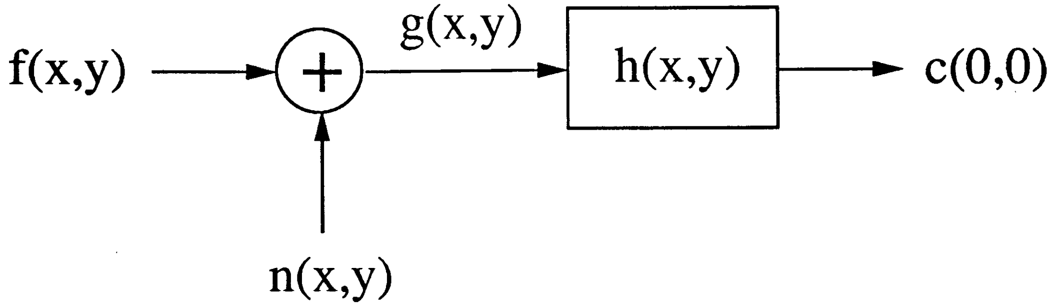


Figure A-1: Object recognition as a signal processing task. $h(x, y)$ is the appropriate linear filter, and $c(0,0)$ is the filter output, or correlation of $h(x, y)$ with $g(x, y)$.

In vector notation, Equation A.1 becomes

$$\begin{aligned} d_{\mathbf{x}}^2 &= (\mathbf{I_x} - \mathbf{k})^t(\mathbf{I_x} - \mathbf{k}) \\ &= \mathbf{I_x}^t\mathbf{I_x} - 2\mathbf{I_x}^t\mathbf{k} + \mathbf{k}^t\mathbf{k} \end{aligned} \tag{A.2}$$

where $\mathbf{I_x}$ is the local N by M subimage whose upper right-hand corner is located at $I(x, y)$, and $d_{\mathbf{x}}$ is $d(x, y)$. The third term of this equation is a constant for a given template $\mathbf{k}$ and may be ignored. If we can assume that local image energy is approximately constant in the scene, then $\mathbf{I_x}^t\mathbf{I_x}$ can also be ignored as a constant. Minimizing $d_{\mathbf{x}}$ then becomes equivalent to maximizing $\mathbf{I_x}^t\mathbf{k}$, which is the cross-correlation of the image and the template. This is equivalent to an inner product, or a projection of the subimage $\mathbf{I_x}$ onto the template $\mathbf{k}$.

This enables object recognition to be posed as a linear signal processing problem where the task is to detect a reference image patch in the presence of noise using a linear filter, as in Figure A-1. The filtering operation is implemented as a correlation[1] to detect the reference object, using the filter $h(x, y)$ as the template. Typically the template is chosen to be an image of the object (or feature) to be recognized. This is called a "matched spatial filter" (MSF), and is well known to be the optimal filter in the case of additive white Gaussian noise [77].

Correlation with an MSF is more convenient as a comparison technique than the

[1] Actually a convolution, but since the correlation filter of $h(x, y)$ is equivalent to the convolution of $h(-x, -y)$, I will omit the distinction.

direct distance measure because it is faster to implement and it is simpler to analyze since it is a linear filter. Because correlation in the spatial domain is equivalent to multiplication in the frequency domain, a large-kernel correlation is much faster when implemented using the Fast Fourier Transform.

A.2 Comparison with eigenfaces

The eigenfaces are used in the recognition system in a manner similar to a set of matched filters. The first step of the distance from face space measure is a set of correlations with the input image using the eigenfaces as the correlation kernels. The outputs of these correlation operations are used to determine the face map $\varepsilon(x, y)$ and the identity by finding the nearest face class Ω_k. Why not just use matched filters directly, where an image of each individual is correlated with the input image, looking for the best match? Is there an advantage to using the eigenfaces?

Consider a face recognition system using M matched filters corresponding to face images of each of the M known people. To recognize a face from this database, there must be M separate correlations performed on the incoming image, resulting in M individual "face maps". Although the filter templates themselves will be highly correlated, the M answers must be treated as unrelated since the correlation among the templates is unknown. Only one template is relevant to the question "Is Joe's face in this image?", and that is Joe's template.

Because the eigenfaces are uncorrelated (they are an orthonormal set), the projection onto each template is meaningful. Rather than looking for the largest output, as in direct convolution, the known relationship among the templates allows for classification based on the ensemble of responses. Joe's presence or absence is indicated by the output of all the filters. Furthermore, the high degree of correlation among the face templates allows fewer than M filters to adequately represent the known faces. As the number of faces increases, relatively fewer new filter kernels are necessary to account for the variation among the face images.

So the eigenfaces in essence implement a "generalized correlation" scheme which is more useful for face classification than straightforward correlation. It is also more efficient, because it can take advantage of the high degree of correlation among the known face images by using only the eigenfaces which have the most discriminating power. In addition, the eigenfaces also allow for a measure of the "faceness" of an image — its distance from face space spanned by the filters — which cannot be done with highly correlated face templates.

A.3 The noise model

The effectiveness of correlation as an approach to object recognition depends on the accuracy and descriptive power of the noise model $n(x, y)$ in Figure A-1. To be useful in general recognition tasks, the noise model must account for all factors which affect the image intensity values captured from a given scene, in particular geometry, surface reflectances, scene illumination, and viewpoint. These factors and their relationships are impossible to model in a linear fashion, and therefore template matching via correlation has a number of shortcomings which make it impractical as a general object recognition scheme. An huge number of templates would have to be created and matched against the image to account for general variations in object appearance.

Normalized correlation [77, 91] solves a small part of the problem because of its insensitivity to amplitude scaling and constant offsets. With normalized correlation the subimage $\mathbf{I_x}$ produces the same correlation output as $a\mathbf{I_x} + \mathbf{b}$, where a and $\mathbf{b}$ are scalar and vector constants, respectively. The output of normalized correlation is unity when the subimage exactly matches the kernel, and less than one otherwise.

Kumar *et al.* [57, 22] extended the deterministic correlator model of Figure A-1 to the case of a stochastic process $f(x, y)$ to determine the best filter for detecting the object in a noisy version of a distorted image. The distortions were defined by a training set of images, and could include transformations such as scale, orientation, changes in illumination, etc. The optimal filter in this case was found to be the principal-component, or first eigenimage. This filter was found to perform much better than conventional matched spatial filters.

The straightforward extension to this approach for face recognition would be to create a filter for each individual, by taking a number of face images of that person under varying conditions and calculating the principal-component of the set. This data set could account for one of the most significant sources of variation in face images, changes in expression. This would still result in a filter kernel per known face, however, and encounter the problems mentioned earlier with this approach. The generation of eigenfaces in its current state is a compromise between the discrimination ability of individual faces from each other and of faces from other classes of objects.

Even with both a simple noise model and the statistical filters based on ensembles of training images described by Kumar *et al.*, correlation techniques are not powerful enough to perform general object recognition. Most computer vision research in the past decade has been devoted to early vision as characterized by Marr [64]:

> The purpose of early visual processing is to sort out which changes are due to what factors and hence to create representations in which the four factors are separated.

As I argue in Chapter 1, however, "visual behaviors" such as face recognition may be to some degree exempt from such an extensive process. Because face recognition is characterized by a limited range of expected views and transformations, representations based on eigenfaces or statistical filters may be sufficient for a useful level of recognition proficiency. A recent correlation-based technique to recognize faces by Burt *et al.* [16, 17, 18] has been demonstrated to work in a limited environment. The performance of the system presented in this thesis, as well as other special-purpose face recognition approaches (e.g. see [8]) support the idea of useful visual behaviors coexisting with more general purpose vision.

Bibliography

[1] H. Abdi, "A generalized approach for connectionist auto-associative memories: interpretation, implication and illustration for face processing," in J. Demongeot, T. Hervé, V. Rialle, and C. Roche (eds.), *Artificial Intelligence and Cognitive Sciences*, Manchester University Press, 1988.

[2] S. Akamatsu, T. Sasaki, H. Fukamachi, and Y. Suenaga, "A robust face identification scheme — KL expansion of an invariant feature space," to appear in *Proc. SPIE Intelligent Robots and Computer Vision X*, Boston, MA, 1991.

[3] J. Aloimonos, "Purposive and qualitative active vision," *Proc. European Conference on Computer Vision*, 1990.

[4] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, Vol. 76, pp. 996-1005, 1988.

[5] R. J. Baron, "Mechanisms of human facial recognition," *Int. J. Man-Machine Studies* **15**, pp. 137-178, 1981.

[6] G. C. Baylis, E. T. Rolls, and C. M. Leonard, "Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey," *Brain Research*, 342, pp. 91-102, 1985.

[7] A. Benton, "Facial recognition 1990," *Cortex*, 26, pp. 491-499, 1990.

[8] M. Bichsel, "Strategies of robust object recognition for the automatic identification of human faces," Ph.D. Thesis, ETH, Zurich, 1991.

[9] W. W. Bledsoe, "The model method in facial recognition," Panoramic Research Inc., Palo Alto, CA, Rep. PRI:15, Aug. 1966.

[10] W. W. Bledsoe, "Man-machine facial recognition," Panoramic Research Inc., Palo Alto, CA, Rep. PRI:22, Aug. 1966.

[11] J. L. Bradshaw and G. Wallace, "Models for the processing and identification of faces," *Perception and Psychophysics*, Vol. 9 (5), pp. 443-448, 1971.

[12] S. E. Brennan, "Caricature generator," Unpublished master's thesis, MIT Media Laboratory, 1982.

[13] R. Brooks, "Intelligence without representation," *Artificial Intelligence* 47, pp. 139-159, 1991.

[14] V. Bruce, *Recognising Faces*, Lawrence Erlbaum Associates, London, 1988.

[15] V. Bruce and A. Young, "Understanding face recognition," *British Journal of Psychology*, *77*, pp. 305-327, 1986.

[16] P. Burt, "Algorithms and architectures for smart sensing," *Proc. Image Understanding Workshop*, pp. 139-153, April 1988.

[17] P. Burt, "Smart sensing within a Pyramid Vision Machine," *Proc. IEEE*, Vol. 76, No. 8, pp. 1006-1015, Aug. 1988.

[18] P. Burt, Personal communication, 1990.

[19] S. R. Cannon, G. W. Jones, R. Campbell, and N. W. Morgan, "A computer vision system for identification of individuals," *Proc. IECON*, Vol. 1, pp. 347-351, 1986.

[20] S. Carey and R. Diamond, "From piecemeal to configurational representation of faces," *Science*, Vol. 195, pp. 312-313, Jan. 21, 1977.

[21] D. Casasent and W.-T. Chang, "Correlation synthetic discriminant functions," *Applied Optics*, Vol. 25, No. 14, pp. 2343-2350, 15 July 1986.

[22] D. Casasent, B. V. K. V. Kumar, and H. Murakami, "A correlator for optimum two-class discrimination," *Electro-Opt. Syst. Des.*, **248**, pp. 321, 1981.

[23] C. S. Choi, T. Okazaki, H. Harashima, and T. Takebe, "Basis generation and description of facial images using principal-component analysis," *Graphics and CAD*, Vol. 46, No. 7, 1990 (in Japanese).

[24] G. W. Cottrell and J. Metcalfe, "EMPATH: Face, gender and emotion recognition using holons," In R.P. Lippman, J. Moody, and D.S. Touretzky (Eds.), *Advances in neural information processing systems 3*, San Mateo, CA: Morgan Kaufmann.

[25] I. Craw, H. Ellis, and J. R. Lishman, "Automatic extraction of face features," *Pattern Recognition Letters* 5, pp. 183-187, 1987.

[26] I. Craw, "Face research in Aberdeen," Mathematical Sciences Dept., Aberdeen, October 2, 1990.

[27] A. R. Damasio, H. Damasio, and G. W. Van Hoesen, "Prosopagnosia: anatomic basis and behavioral mechanisms," *Neurology* Vol. 32, pp. 331-41, April 1982.

[28] C. Darwin, *The Expression of the Emotions in Man and Animals*, London: John Murray, 1872.

[29] G. M. Davies, H. D. Ellis, and J. W. Shepherd (eds.), *Perceiving and Remembering Faces*, Academic Press, London, 1981.

[30] E. H. F. de Haan, A. Young, and F. Newcombe, "Face recognition without awareness," *Cognitive Neuropsychology*, *4* (4), pp. 385-415, 1987.

[31] R. Desimone, T. D. Albright, C. G. Gross, and C. J. Bruce, "Stimulus-selective properties of inferior temporal neurons in the macaque," *Neuroscience* 4, pp. 2051-2068, 1984.

[32] R. Desimone, "Face-selective cells in the temporal cortex of monkeys," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 1-8, 1991.

[33] R. Diamond and S. Carey, "Why faces are and are not special: an effect of expertise," *J. Exp. Psych: G*, Vol. 115, No. 2, pp. 107-117, 1986.

[34] M. J. Farah, *Visual Agnosia*, MIT Press, Cambridge, MA, 1990.

[35] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Computers*, Vol. c-22, No. 1, pp. 67-92, January 1973.

[36] M. Fleming and G. Cottrell, "Categorization of faces using unsupervised feature extraction," *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2, pp. 65-70, San Diego, CA, June 1990.

[37] H. D. Ellis and M. Florence, "Bodamer's (1947) paper on prosopagnosia," *Cognitive Neuropsychology*, *7* (2), pp. 81-105, 1990.

[38] N. L. Etcoff, R. Freeman, and K. R. Cave, "Can we lose memories of Faces? Content specificity and awareness in a prosopagnosic," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 25-41, 1991.

[39] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Identification of human faces," *Proc. IEEE*, Vol. 59, pp. 748-760, 1971.

[40] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Man-machine interaction in human-face identification," *The Bell System Technical Journal*, Vol. 41, No. 2, pp. 399-427, Feb. 1972.

[41] C. C. Goren, M. Sarty, and R. W. K. Wu, "Visual following and pattern discrimination of face-like stimuli by newborn infants," *Pediatrics, 56*, pp. 544-549, 1975.

[42] G. Gordon, "Face recognition based on depth maps and surface curvature," *Geometric Methods in Computer Vision*, Proc. SPIE 1570, 1991.

[43] V. Govindaraju, D. B. Sher, R. K. Srihari, and S. N. Srihari, "Locating human faces in newspaper photographs," *Proc. CVPR*, San Diego, CA, pp. 549-554, June 1989.

[44] L. D. Harmon, "Some aspects of recognition of human faces," in O. J. Grusser and R. Klinke (eds.), *Pattern Recognition in Biological and Technical Systems* Springer-Verlag, Berlin, 1971.

[45] L. D. Harmon and W. F. Hunt, "Automatic recognition of human face profiles," *Computer Graphics and Image Processing*, **6**, pp. 135-156, 1977.

[46] L. D. Harmon, M. K. Khan, R. Lasch, and P. F. Ramig, "Machine identification of human faces," *Pattern Recognition*, Vol. 13, No. 2, pp. 97-110, 1981.

[47] M. H. Harries and D. I. Perrett, "Visual processing of faces in temporal cortex: physiological evidence for a modular organization and possible anatomical correlates," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 9-24, 1991.

[48] D. C. Hay and A. W. Young, "The human face," in A. W. Ellis (ed.), *Normality and Pathology in Cognitive Functions*, Academic Press, London, 1982.

[49] E. I. Hines and R. A. Hutchinson, "Application of multi-layer perceptrons to facial feature location," IEE *Third International Conference on Image Processing and Its Applications*, pp. 39-43, July 1989.

[50] T. Kanade, "Picture processing system by computer complex and recognition of human faces," Dept. of Information Science, Kyoto University, Nov. 1973.

[51] M. D. Kelly, "Visual identification of people by computer," Stanford Artificial Intelligence Project Memo AI-130, July 1970.

[52] K. M. Kendrick and B. A. Baldwin, *Science* 236, pp. 448-450, 1987.

[53] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 1, pp. 103-108, Jan. 1990.

[54] J. Kittler and P. C. Young, "A new approach to feature selection based on the Karhunen-Loeve expansion," *Pattern Recognition*, Vol. 5, pp. 335-352, 1973.

[55] T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, 1989.

[56] T. Kohonen, E. Oja, and P. Lehtio, "Storage and processing of information in distributed associative memory systems," in G. E. Hinton and J. A. Anderson (eds.), *Parallel Models of Associative Memory*, Hillsdale, NJ: Lawrence Erlbaum Associates, 105-143, 1981.

[57] B. V. K. V. Kumar, D. Casasent, and H. Murakami, "Principal-component imagery for statistical pattern recognition correlators," *Optical Engineering*, Vol. 21 No. 1, pp. 43-47, Jan/Feb 1982.

[58] J. T. Lapresté, J. Y. Cartoux, and M. Richetin, "Face recognition from range data by structural analysis," in G. Ferraté et al. (eds.), *Syntactic and Structural Pattern Recognition*, NATO ASI Series, Vol. F45, Springer-Verlag, 1988.

[59] K. R. Laughery and M. S. Wogalter, "Forensic applications of facial memory research," in A. W. Young and H. D. Ellis (eds.), *Handbook of Research on Face Processing*, Elsevier Science Publishers B. V. (North-Holland), 1989.

[60] J. C. Lee and E. Milios, "Matching range images of human faces," *Proc. ICCV*, Osaka, Japan, pp. 722-726, Dec. 1990.

[61] S. Leehey, "Face recognition in children: evidence for the development of right hemisphere specialization," Ph.D. Thesis, Dept. of Psychology, Massachusetts Institute of Technology, May 1976.

[62] J. Liggett, *The Human Face*, Stein and Day, New York, 1974.

[63] D. Lowe, "Three-dimensional object recognition from single two-dimensional images," Technical Report No. 202, Courant Institute, Feb. 1986.

[64] D. Marr, *Vision*, W. H. Freeman, San Francisco, 1982.

[65] H. Midorikawa, "The face pattern identification by back-propagation learning procedure," *Abstracts of the First Annual INNS Meeting*, Boston, p. 515, 1988.

[66] M. Minsky, *The Society of Mind*, Simon and Schuster, New York, 1986.

[67] R. C. Nelson, "Qualitative detection of motion by a moving observer," *Proc. CVPR*, Maui, Hawaii, pp. 173-178, June 1991.

[68] "Faces From the Future," *Newsweek*, Feb. 13, 1989, p. 62.

[69] A. J. O'Toole, R. B. Millward, and J. A. Anderson, "A physical system approach to recognition memory for spatially transformed faces," *Neural Networks*, Vol. 1, pp. 179-199, 1988.

[70] A. J. O'Toole and H. Abdi, "Connectionist approaches to visually-based facial feature extraction," in G. Tiberghien (ed.), *Advances in Cognitive Science*, Vol. 2, Ellis Horwood Ltd., 1989.

[71] S. E. Palmer, "The psychology of perceptual organization: a transformational approach," in J. Beck, B. Hope, and A. Rosenfeld (eds.), *Human and Machine Vision*, Academic Press, New York, 1983.

[72] D. I. Perrett, E. T. Rolls, and W. Caan, "Visual neurones responsive to faces in the monkey temporal cortex," *Exp Brain Res*, Vol. 47, pp. 329-342, 1982.

[73] D. I. Perrett, P. A. J. Smith, D. D. Potter, A. J. Mistlin, A. S. Head, A. D. Milner, and M. A. Jeeves, "Visual cells in the temporal cortex sensitive to face

view and gaze direction," *Proceedings of the Royal Society of London, 223*, pp. 293-317, 1985.

[74] D. I. Perrett, A. J. Mistlin, and A. J. Chitty, "Visual neurones responsive to faces," *TINS*, Vol. 10, No. 9, pp. 358-364, 1987.

[75] D. I. Perrett, A. J. Mistlin, A. J. Chitty, P. A. J. Smith, D. D. Potter, R. Broennimann, and M. Harries, "Specialized face processing and hemispheric asymmetry in man and monkey: evidence from single unit and reaction time studies," *Behavioural Brain Research*, 29, pp. 245-258, 1988.

[76] J. B. Pittenger and R. E. Shaw, "Ageing faces as viscal-elastic events: Implications for a theory of nonrigid shape perception," *Journal of Experimental Psychology: Human Perception and Performance, 1*, pp. 374-382, 1975.

[77] W. K. Pratt, *Digital Image Processing*, John Wiley & Sons, New York, 1978.

[78] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Detection of interest points using symmetry," *Proc. ICCV*, Osaka, Japan, pp. 62-65, Dec. 1990.

[79] G. Rhodes, S. Brennan, and S. Carey, "Identification and ratings of caricatures: implications for mental representations of faces," *Cognitive Psychology* **19**, pp. 473-497, 1987.

[80] M. J. Riddoch and G. W. Humphreys, "A case of integrative visual agnosia," *Brain*, **110**, pp. 1431-1462, 1987.

[81] E. T. Rolls and G. C. Baylis, "Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey," *Exp Brain Res*, Vol. 65, pp. 38-48, 1986.

[82] E. T. Rolls, G. C. Baylis, M. E. Hasselmo, and V. Nalwa, "The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey," *Exp Brain Res*, Vol. 76, pp. 153-164, 1989.

[83] O. Sacks, *The Man Who Mistook His Wife For a Hat*, Harper & Row, New York, 1987.

[84] T. Sakaguchi, O. Nakamura, and T. Minami, "Personal identification through facial images using isodensity lines," SPIE Vol 1199 *Visual Communications and Image Processing IV*, pp. 643-654, 1989.

[85] T. Sakai, M. Nagao, and M. Kidode, "Processing of multilevel pictures by computer — the case of photographs of human faces," *Systems, Computers, Controls*, Vol. 2, No. 3, pp. 47-53, 1971.

[86] Y. Satoh, Y. Miyake, H. Yaguchi, and S. Shinohara, "Facial pattern detection and color correction from negative color film," *Journal of Imaging Technology*, Vol. 16, No. 2, pp. 80-84, April 1990.

[87] S. Sclaroff and A. Pentland, "Closed-form solutions for physically-based shape modeling and recognition," *Proc. CVPR*, Maui, Hawaii, pp. 238-243, June 1991.

[88] J. Sergent and J.-G. Villemure, "Prosopagnosia in a right hemispherectomized patient," *Brain*, **112**, pp. 975-995, 1989.

[89] M. A. Shackleton and W. J. Welsh, "Classification of facial features for recognition," *Proc. CVPR*, Maui, Hawaii, pp. 573-579, June 1991.

[90] T. Shallice and M. Jackson, "Lissauer on agnosia," *Cognitive Neuropsychology*, *5* (2), pp. 153-192, 1988.

[91] W. M. Silver, "Alignment and gauging using normalized correlation search," Technical Report, Cognex Corporation, Needham, Mass.

[92] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, Vol. 4, No. 3, pp. 519-524, March 1987.

[93] M. A. Sokolov, "Visual motion: algorithms for analysis and application," Vision and Modeling Group Technical Report #127, MIT Media Lab, February 1990.

[94] T. J. Stonham, "Practical face recognition and verification with WISARD," in H. Ellis, M. Jeeves, F. Newcombe, and A. Young (eds.), *Aspects of Face Processing*, Martinus Nijhoff Publishers, Dordrecht, 1986.

[95] G. Strang, *Linear Algebra and Its Applications*, Harcourt Brace Jovanovich, San Diego, 1988.

[96] M. J. Swain, "Color indexing," Technical Report 360, Computer Science Dept., University of Rochester, November 1990.

[97] D. Tock, I. Craw, and R. Lishman, "A knowledge based system for measuring faces," Mathematical Sciences Department, University of Aberdeen, Scotland.

[98] D. Tranel and A. R. Damasio, "Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics," *Science*, Vol. 228, pp. 1453-1454, 21 June 1985.

[99] M. Turk and A. Pentland, "Face processing: models for recognition," SPIE Vol. 1192, *Intelligent Robots and Computer Vision VIII*, 1989.

[100] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, 1991.

[101] M. Turk and A. Pentland, "Face recognition using eigenfaces," *Proc. CVPR*, Maui, Hawaii, pp. 586-591, June 1991.

[102] S. Ullman, "An approach to object recognition: aligning pictorial descriptions," AI Memo No. 931, MIT Artificial Intelligence Laboratory, Dec. 1986.

[103] S. Watanabe, *Pattern Recognition: Human and Mechanical*, Wiley-Interscience, New York, 1985.

[104] K. Wong, H. Law, and P. Tsang, "A system for recognising human faces," *Proc. ICASSP*, pp. 1638-1642, May 1989.

[105] S. Yamane, S. Kaji, and K. Kawano, "What facial features activate face neurons in the inferotemporal cortex of the monkey?" *Exp Brain Res*, Vol. 73, pp. 209-214, 1988.

[106] R. K. Yin, "Looking at upside-down faces," *J. Exp. Psychol.*, 81, pp. 141-145, 1969.

[107] A. W. Young, K. H. McWeeny, D. C. Hay, and A. W. Ellis, "Matching familiar and unfamiliar faces on identity and expression," *Psychol Res*, Vol. 48, pp. 63-68, 1986.

[108] A. L. Yuille, D. S. Cohen, and P. W. Hallinan, "Feature extraction from faces using deformable templates," *Proc. CVPR*, San Diego, CA, pp. 104-109, June 1989.

[109] A. L. Yuille, "Deformable templates for face recognition," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 59-70, 1991.