DISSERTATION

AN ALGORITHMIC IMPLEMENTATION OF EXPERT OBJECT RECOGNITION IN VENTRAL VISUAL PATHWAY

Submitted by

Kyungim Baek

Department of Computer Science

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2002

COLORADO STATE UNIVERSITY

August, 2002

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY KYUNGIM BAEK ENTITLED AN ALGORITHMIC IMPLEMENTATION OF EXPERT OBJECT RECOGNITION IN VENTRAL VISUAL PATHWAY BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Committee Member

Committee Member

Committee Member

Adviser

Department Head

ABSTRACT OF DISSERTATION

AN ALGORITHMIC IMPLEMENTATION OF EXPERT OBJECT RECOGNITION IN VENTRAL VISUAL PATHWAY

Understanding the mechanisms underlying visual object recognition has been an important subject in both human and machine vision since the early days of cognitive science. Current state-of-the-art machine vision systems can perform only rudimentary tasks in highly constrained situations compared to the powerful and flexible recognition abilities of the human visual system.

In this work, we provide an algorithmic analysis of psychological and anatomical models of the ventral visual pathway, more specifically the pathway that is responsible for expert object recognition, using the current state of machine vision technology. As a result, we propose a biologically plausible expert object recognition system composed of a set of distinct component subsystems performing feature extraction and pattern matching.

The proposed system is evaluated on four different multi-class data sets, comparing the performance of the system as a whole to the performance of its component subsystems alone. The results show that the system matches the performance of state-of-the-art machine vision techniques on uncompressed data, and performs better when the stored data is highly compressed.

Our work on building an artificial vision system based on biological models and theories not only provides a baseline for building more complex, end-to-end vision

systems, but also facilitates interactions between computational and biological vision studies by providing feedback to both communities.

Kyungim Baek
Department of Computer Science
Colorado State University
Fort Collins, Colorado 80523
Fall 2002

ACKNOWLEDGEMENTS

This work would not be possible without help and support of many people. First I would like to thank my adviser, Dr. Bruce Draper, for his continued guidance, inspiration and support, for sharing many of his insights, and for uncountable discussions and revisions of this dissertation. I owe much to him for the freedom he gave me to pursue my own interests. I am grateful to my committee members, Dr. Charles Anderson, Dr. Ross Beveridge, and Dr. Michael Kirby for their suggestions and advice. I want to thank Jeff Boody and Jeremy Hayes for the help running some experiments, Jose Bins, Emanuel Grant, and Sohyun Kown for all their support and encouragement that kept me going.

A special thanks to my family for their love and support over the years.

Dedicated to my parents,

who gave me

everything

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

How do humans identify and classify objects? This simple question has formed an active area of study in both human and machine vision. As we experience in every moment of our life, the human visual system exhibits an amazing capability to recognize objects. People know about a great number of different types of objects, yet they can identify the object in front of them almost effortlessly under widely varying circumstances such as changes in viewing position, illumination, occlusion, and object shape. Current state-of-the-art machine vision systems, however, can perform only rudimentary tasks in highly constrained situations and, therefore, their recognition abilities are far less powerful and flexible than the capability of the human visual system.

There are many factors that make building an artificial object recognition system a difficult task. We have only a poor understanding of the mechanisms underlying the recognition process. When we see 3D objects in a scene we receive 2D stimulation on our retina, which is transformed into neural signals. Then, the visual information (signal) is sent to the brain over multiple pathways through different cortical areas, each of which processes the data until a final decision about the objects' identities is made. The problem is that we do not know how the visual processes are performed, how the inputs and outputs of each process are characterized, in what forms and how we store our understanding or knowledge about objects from past experience, or how we extract information from our memory to make decisions. All of these questions boil down to the previously posited, more comprehensive question: "How does the human brain solve the visual object recognition problem?"

The question has been a topic of study since the early days of cognitive science. Scientists in the fields of psychophysics, psychology, neuroscience, cognitive neuroscience, and computer science have made tremendous efforts to understand the mechanisms underlying visual perception and theorize computational models for building artificial vision systems. Research in these areas not only enriches our knowledge of visual perception – we now have an understanding of many visual phenomena, anatomical structures of visual areas in brain, and functional features related to some of those areas – but also provides a large number of theories and models that have been continuously explored and revised.

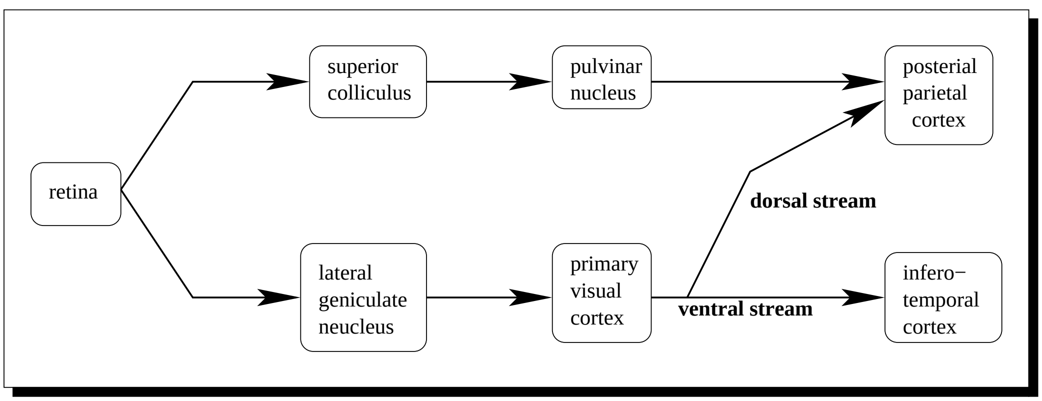


Figure 1.1: Diagram of the major routes of visual processing in the primate visual system [102].

Figure 1.1 shows a result of such efforts. It illustrates the major routes of visual processing in the primate visual system. Information about a scene is captured by the photoreceptors in the retina which convert light into electrical signals. The signals generated by photorecptors are transmitted to the *lateral geniculate nucleus* (LGN) and the *superior colliculus* (SC) of the midbrain through the optic nerve connected to the retinal ganglion cells. Visual information processed in the SC is conveyed to the pulvinar nucleus of the thalamus, and eventually arrives at the posterial parietal cortex. Traditionally, this route is interpreted to be responsible for saccadic eye movements. Visual information in LGN is further projected onto the primary visual cortex, where the two major cortical streams originate. The ventral stream, which ends in the infero-temporal cortex, is known to be responsible for visual perception, while the dorsal stream is considered as a visuo-motor pathway, which runs dorsally to the posterior parietal cortex.

Theories and findings in visual neuroscience have been applied to the design of innovative algorithms for computer vision, and some of the most successful computer vision algorithms have direct biological inspirations [3, 91, 96, 97]. Although they have provided many useful applications, these previous attempts focused on models of early vision such as edge detection and color analysis, or partial computational elements that are roughly in the dorsal visual pathway such as motion detection, 3D surface reconstruction, and perceptual organization. However, since the two broad cortical pathways were found in the monkey by Ungerleider and Mishkin [104], it has been generally considered that the ventral pathway plays the critical role in identifying and recognizing objects.

In this work, we have tried to provide a possible algorithmic analysis of psychological and anatomical models of the ventral visual pathway, more specifically the pathway within the ventral stream that is responsible for recognizing familiar objects seen from familiar viewpoints, using the current state of machine vision technologies. This work is mainly inspired by two biological theories: Stephen Kosslyn's psychophysical model of visual perception [82] and Michael Tarr and his colleagues' work on a viewpoint-dependent mechanism and perceptual expertise for visual object recognition [49, 51, 144, 145].

The overall structure of our approach is based on Kosslyn's model of visual object recognition, in which the ventral visual pathway is composed of a set of functionally distinctive and anatomically localized components that interact with each other. However, while Kosslyn's model provides a good starting point to build practical artificial vision systems that are biologically inspired, he concentrates more on how the boundaries that delimit distinct processing subsystems are specified than how the subsystems achieve their computational goals. As described in Chapter 3, there has been a debate on computational mechanisms for visual object recognition in the brain. Recent work by Tarr and his colleagues has shown converging behavioral and psychological evidence for viewpoint-dependent mechanisms for visual perception, which provide strong support for viewpoint-dependent, appearance-based methods for object recognition in the machine vision community. Based on Kosslyn's model and Tarr's theory, we constructed a more complete end-to-end object recognition system in which a set of interacting yet relatively independent subsystems implement each of the components.

We begin this introduction with a brief overview of biological vision systems in which visual processing in the primate is characterized by functional specialization from the very beginning. The distinctive functionality is related to the two cortical visual pathways: dorsal and ventral pathway. We then provide a general description of Kosslyn's psychophysical model of high-level visual processing, in which the primate vision system consists of multiple processing subsystems interacting with each other rather than one single process. Then, we describe the proposed approach for building a computational analogue to the specialized part of ventral visual pathway for recognizing familiar objects seen from familiar viewpoints. We conclude this introductory chapter by sketching an outline of the rest of the thesis.

1.1 Biological Vision System

For humans, the sense of vision is a dominant sense, playing a central role in our interaction with the environment. Standard accounts of vision implicitly assume that the purpose of the visual system of an organism is to obtain knowledge of its surroundings so as to behave appropriately and in accordance with its current behavioral goals. From this perspective, the success of the visual process requires that some form of object identification and movement detection take place based on size, shape, color, location, and past experience.

A central principle that characterizes vision is functional specialization. Specialization in vision occurs at the very earliest point possible: in the photoreceptors. There are four different types of photoreceptors that are grouped into two classes: the rod and S-, M-, and L-type cone photoreceptors [99]. The rods are much more sensitive to low levels of illumination than the cones; the cones are tuned to specific color bands. The functional specialization continues as the optic nerve, a bundle of fibers, carries visual information from the eyes to the brain. The magnocellular fibers tend to favor information that varies temporally, such as motion or flicker, while parvocellular fibers tend to carry information about static properties such as color, orientation, or depth [102]. These fibers are connected to the LGN, and the visual information is transfered to the first visual area in the cortex (area V1, which is also known as striate cortex, primary visual cortex, or area 17) through two LGN channels - the parvo and magno channels.[1]

Anatomically, the early visual cortex is divided into five separate areas: V1 to V5. As described above, V1 receives visual information directly from the LGN. Since the

[1]In 1994, a third channel from LGN to V1 was found by Hendry and Yoshioka [62]. However, its role has not been clearly identified.

ground-breaking discovery of orientation selectivity in V1 cells by Hubel and Wiesel [64], mountains of information on V1 has been accumulated. It has been shown that, in addition to orientation, there are cells in V1 that are selective for other properties, such as direction of motion, wave-length, and the length of a bar-type stimulus. V1 seems to make those features explicit and provide them as input to other cortical areas for further processing.

Compared to V1, other areas of the cortex remain a relatively wild neuroscientific frontier. However, recent advances of technologies measuring brain activities, such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and repetitive transcranial magnetic stimulation (rTMS), provide data for modeling higher level visual processing in brain. Results from various areas of cognitive science based in part on the new technologies suggest that different cortical regions appear to be dedicated to different visual attributes. For example, V2 seems specialized to process form information, which would be helpful for figure-ground separation and object shape identification. Cells in V3 are selective for orientation, and many are also tuned to motion and to depth although the cell properties provide few clues to the function of V3 [102]. Also, it has been postulated that V4 is involved in color perception and V5, also known as the middle temporal (MT) area, processes motion and depth information [154, 160].

This functional specialization is intimately related to two large-scale cortical pathways of visual processing, one originating from the primary visual cortex projecting ventrally to the inferior temporal (IT) cortex and the other projecting dorsally to the posterior parietal (PP) cortex (Figure 1.2). Historically, these two distinct streams are also known as the "what" and "where" pathways based on their role in visual processing – object identification vs. object localization [104]. The existence of such distinct pathways has been generally accepted based upon considerable evidence from animal and human studies [82, 83]. Milner and Goodale, however, view this functional

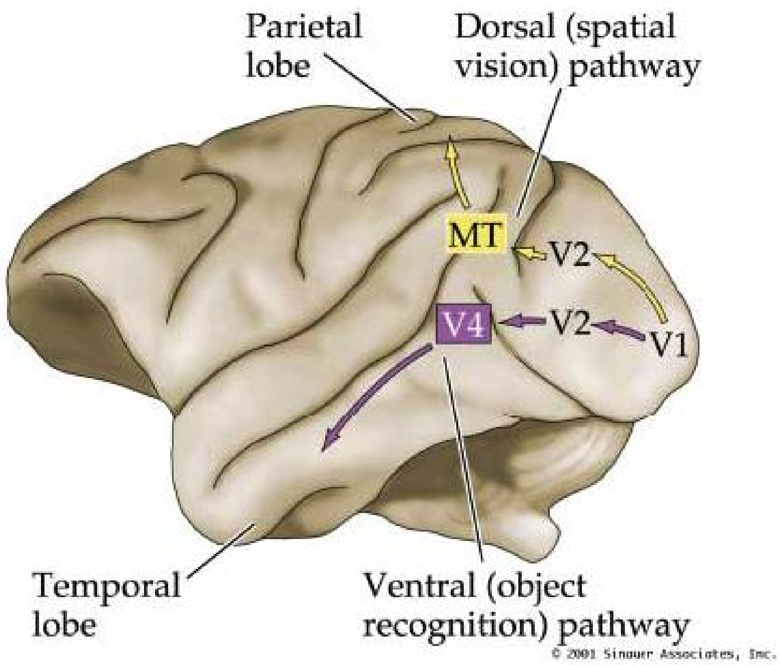


Figure 1.2: The two visual processing pathways in the primate cerebral cortex (reprinted from [162]).

distinction with a somewhat different perspective: instead of the subdomains of perception, they describe the different role of the two pathways as perception and visually guided action [102]. In this perspective, the dorsal pathway is responsible for vision in support of immediate physical action and, therefore, models the world in egocentric coordinates with virtually no memory. On the other hand, the ventral pathway is responsible for visual perception and maintains visual memory for allocentric modeling of objects in the environment.

Milner and Goodale further show that multiple subpathways may exist within the two broad pathways. For example, the dorsal pathway can be further divided into anatomically distinct components for different egocentric coordinates, such as eye-centered, head-centered, and shoulder-centered subsystems [102]. Recent neuroimageing studies showing preferential activity patterns in discrete areas of ventral pathway to different objects – faces, houses, chairs, and places – also support the hypothesis that multiple subpathways exist in the ventral stream [71, 112]. The intensive brain imaging studies on face recognition, in particular, lead to a debate on a face specific pathway; "Is it really specialized for faces only or the objects for which people have developed expertise [33, 49, 51, 76, 125, 145]?" Although it has not yet been resolved completely, more recent results that combine behavioral, psychological and brain-imaging studies seem to suggest that the pathway is more likely to be an expert object recognition subsystem rather than a specialized face recognizer.

1.2 Kosslyn's Psychophysical Model of Visual Perception

The studies of biological vision systems described in Section 1.1 have helped us to understand functional roles and anatomical structures of visual areas in the brain. Although they have provided much information about visual perception, these studies do not explain how the bits and pieces can be connected and interact with each other to achieve the perceptual goal. Now we are in need of a psychological and structured model which systematically puts a decades' worth of work together, and that is why we turn to Kosslyn's model of visual perception.

Kosslyn has studied, for at least twenty years, the brain mechanisms underlying visual mental imagery as well as object recognition. His publication, *'Image and Brain'* [82], integrates research on the nature of high-level vision and mental imagery, and provides a computational theory of processing that underlies object recognition and imagery. His theory is based on the idea that visual perception and mental imagery (representation) share common mechanisms, and that mental imagery events in the brain are generated, interpreted, and actually used in perception [82, 83].

Figure 1.3 shows Kosslyn's model of visual object identification, which consists of seven major components. Each component has distinct functionality and is implemented in a separate, relatively small region of the brain [82, 83, 84]. The stimulus input from the eyes generates an image in a structure called the "visual buffer", which corresponds to a set of retinotopically mapped areas in the occipital lobe. To select information for additional processing in the system, an attention window extracts a region of the visual buffer. The information in the attention window is then sent downstream to two major cortical pathways from the occipital lobe; the object properties encoding system that runs ventrally to the inferior temporal lobe, and spatial properties encoding system that runs dorsally to the posterior parietal lobe.

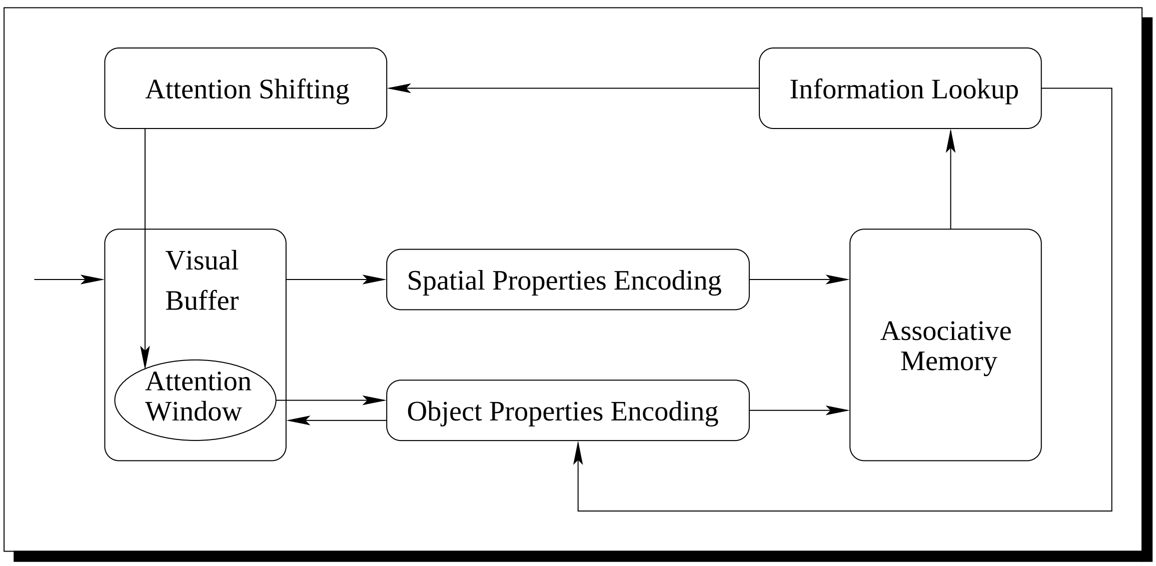


Figure 1.3: Kosslyn's psychophysical model of visual object identification with seven processing components [84].

The ventral stream deals with object properties such as shape, color and texture. The system first extracts features that describe object properties from the input passed from the attention window and then matches those features to representations stored in visual memory. While the goal of the ventral stream is to match and thereby recognize objects, the dorsal stream is mainly responsible for guiding actions (e.g. eye movement) by registering spatial information, such as location, size and orientation of objects or object parts.

The output from the ventral and dorsal systems converge at an associative memory which is a cortical long-term storage structure located partly in the posterior superior temporal cortex. Associative memory stores multimodal information, containing not only perceptual information, but also more abstract conceptual information. If the incoming information is strongly matched with the representation of an object in the associative memory, the object is identified and more knowledge about the object is accessed. However, if the match is not strong enough, object identity is hypothesized and additional information is collected by the information lookup system (located in the dorsolateral prefrontal cortex). Unlike previously discussed architectures, this model has a strong top-down component. The hypotheses about an object's identity guides the search for additional properties that help to determine the presence of the hypothesized object. This bottom-up and then top-down processing mechanism is in fact similar to Lowe's model [92, 93] discussed in the next chapter.

Finally, the attention is shifted if the search process finds a location of informative or distinctive characteristics in the visual buffer. Then, the new attended region is encoded and matched through the ventral and dorsal systems. The object and spatial properties are registered in the associative memory and possibly activate a different representation of the same or different object. The identification process is then applied again.

Kosslyn's model described in this section is for visual object identification in general, which covers not only visual processing, but also intelligence, motor control, and complex object and environmental models. His model, however, makes a strong distinction between the strictly visual system – visual buffer, spatial and object properties encoding systems – and other mixed modality systems – associative memory, information lookup, and attention shifting systems. Our work applies the ventral stream of Kosslyn's model in the more limited context of familiar object recognition seen from familiar viewpoints. A more detailed description of Kosslyn's model of the ventral visual pathway is given in Chapter 3.

1.3 The Proposed System

1.3.1 Introduction

Compared to the capability of the human vision system to recognize objects, current artificial vision systems can perform only rudimentary tasks in highly constrained situations. Thus, researchers have tried to augment studies of biological vision and apply them to designing innovative computer vision algorithms. As a result, interdisciplinary research in computational and psychophysical aspects of object recognition is a very active area of study. The research includes experimental studies of human recognition abilities, computational modeling of the results and the design of practical computer vision systems. Attempts to build complete, biologically inspired vision systems have been rare, however. One of the reasons is that integrated work on biological vision with specifications clear and detailed enough to implement computational models is hard to find.

The primary motivation for this work comes from Kosslyn's functional and psychopysical model of brain mechanisms underlying object recognition [82], and the recently developed theories on the existence of an expert object recognition pathway within the ventral visual stream [51, 145]. As described in Section 1.2, Kosslyn breaks down the recognition process into component subsystems. Each subsystem is anatomically localized in the brain, has distinctive functionality, and interacts with other subsystems to achieve the recognition goal. Therefore, it provides a good structural framework for biologically plausible object recognition systems. Our work follows the principle of "start from small" based on the expert object recognition theory, applying Kosslyn's general model of the ventral visual pathway in the more limited context of recognizing familiar objects from common viewpoints.

1.3.2 System Description

The goal of this work is to reconsider how we design artificial object recognition systems of practical use to more closely mimic biological ones and provide a possible algorithmic analysis using the current state of machine vision technologies. There are plenty of techniques in the field of computer vision that can implement components of biological vision systems. In this work, the ventral visual stream of Kosslyn's psychological model is mapped onto computational algorithms and the resulting system is tested in the context of expert object recognition.

Our approach for building a computational expert object recognition system is illustrated in Figure 1.4. The function of the system is to match the current stimulus image to previously seen images stored in the visual memory. It does not build a 3D model and does not assign a symbolic or linguistic label to the input image, which may include multi-modal information processed beyond the ventral stream. Instead, the system retrieves visually similar images from the memory.

In Figure 1.4, the system consists of two phases – training and run-time (or testing). The input to the system is a set of small image patches, which are assumed to be focused, scaled, rotated, and registered images of target objects. This is what the attention window produces in Kosslyn's model, and we do not directly model the attention mechanism itself in this work.

The visual buffer includes V1, and it is well known that the receptive field profiles of the simple cells in V1 can be approximated reasonably well by Gabor filters, and the complex cells approximate frequency energy functions [121]. Thus, a bank of multi-scale, orientation-selective Gabor filters are applied to the input images to model the operation in V1. The parameters for Gabor functions, such as spatial aspect ratio, spatial frequency bandwidth, and phase offset, are tuned as suggested from studies on biological visual systems [116]. The output generated by the filtering operation are transformed versions of retinal image patches that form an image pyra-

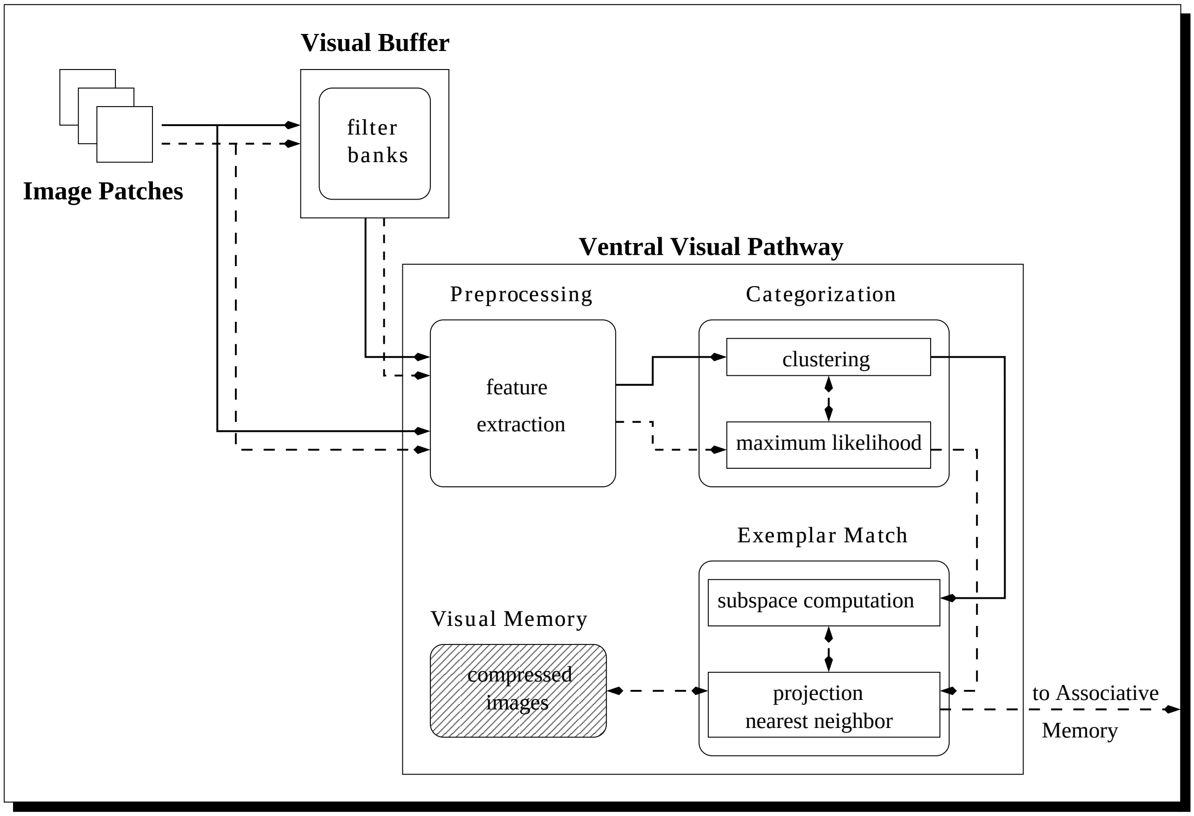


Figure 1.4: Overview of proposed expert object recognition system. The solid line follows training phase, while the dotted line shows the run-time execution.

mid. The operation is basically an image-to-image transformation, so the output is still retinotopic, which is consistent with the architecture of area V1. Both the raw images and filter responses are passed to the preprocessing subsystem where more complex features are extracted.

The pattern matching in the proposed system consists of two separate processes that are responsible for different levels of recognition, referred as *categorical* and *subordinate* levels. The categorization subsystem is responsible for the categorical level recognition. During training, the categorization subsystem is modeled by unsupervised clustering algorithms which group images that are visually similar. Therefore, images in a cluster do not necessarily share semantic properties. In this work, the popular K-Means clustering algorithm [42], Expectation-Maximization (EM [35]) clustering algorithm, and clustering based on local probabilistic PCA are implemented and tested.

The subordinate or instance level recognition is performed by subspace projection and nearest neighbor matching. There are three different unsupervised subspace projection algorithms considered in this work: principal component analysis (PCA [79, 151]), independent component analysis (ICA [9, 34]), and factor analysis (FA [139]). This approach of modeling the exemplar subsystem as subspace projection and matching is our interpretation of Kosslyn's description of visual memory as "compressed images", that do not have topography but contain enough information to reconstruct the original raw images [82].

In Figure 1.4, the dotted arrows show the run-time execution path. For a given input image, a set of filter responses are computed by applying the bank of Gabor filters with different orientation selectivity, phase shift, and multiple scales. Then, a class label is assigned by performing maximum likelihood classification between the input data and each of the clusters. After this categorical level of recognition, the input data is encoded as a compressed image using the labeled cluster's basis vectors computed in the training phase, and the nearest neighbor match retrieves the closely matched instances.

There are two things to note about the system. First, the run-time processing is very fast, which is a property found in human expert recognition system. Second, computing the unique subspace for each cluster formed in the categorization subsystem realizes a local linear subspace approach. It is unlikely that the images are drawn from a single global normal distribution as assumed by global linear models, especially when the objects are from multiple classes. The expert object recognition system tends to deal with many classes of objects. The proposed system has been tested in multi-class domains.

1.3.3 Contributions

The main contribution of this work is a system that implements a psychophysical model of expert visual object recognition supported by evidence from many related fields of study. This system provides explicit connections between computational and biological models of visual object recognition. Many of the vision theories and systems developed previously also have biological relevance, but none of them model human expert object recognition as an end-to-end process, in which every component is based on a biological model. Kosslyn's model describes visual perception by breaking the entire process into components according to their functionality and anatomical localization in the brain. The proposed system strictly follows the structure and data flow depicted in Kosslyn's model and provides each component with a possible algorithmic mapping based on current work on machine vision technology.

Having different levels of recognition in one framework also allows us to build a complete vision system that models Tarr and his colleagues' argument for a single, highly plastic expert visual recognition system [144]. They argue that, for a given task, a single system can adapt itself for different levels of classification, and that this is one of the defining characteristics of expert recognition [54]. A neural network model that accounts for this ability has been developed [156], but our approach shows it in the context of a more complete vision system.

There are many machine vision algorithms that can implement the functionality of the subsystems in the biological model. Therefore, computational choices have to be made among them. In the course of developing a biologically plausible vision system, this work also provides comparative evaluations and performance analysis among algorithms that have the same gross functionality. Apart from the biological relevance, it gives valuable information to the machine vision community.

Our interpretation of the biological models and theories for building an artificial vision system is quite simple. As a result, the proposed system is in its early stages at this moment. This effort, however, provides a baseline for building a more complex, end-to-end vision system based on functional model of biological vision systems. It will benefit both computational and biological vision studies; if the system turns out to be successful, it provides a practical object recognition system that is biologically inspired. Otherwise, we can give valuable feedback to the psychological community about ambiguity, incompatibility with computational techniques, and difficulties in fitting algorithms to their psychological models. This feedback can reduce the gap between computational and theoretical fields of studies and, therefore, facilitate the realization of machine vision system close to that of humans.

1.4 Outline of Thesis

Chapter 2 reviews the computational approaches for visual object recognition in studies from both the computer vision and biological vision literatures. Chapter 3 describes Kosslyn's model of the ventral visual pathway in more detail, and also provides arguments for the existence of an expert object recognition pathway within the ventral stream. The components of the proposed system are described in Chapter 4 and Chapter 5. In these chapters, the possible computational algorithms for implementing each of the subsystems are described in connection with the biological motivation. Chapter 6 shows the results of running the complete system, including evaluation of the effectiveness of the proposed system design. In the course of developing the proposed system, we performed comparative evaluation on several subspace projection algorithms to make a computational choice for implementing the exemplar match subsystem. In Chapter 7, we present these supplementary studies, performed outside the context of proposed system. Finally, in Chapter 8, we give a summary of the thesis, present our conclusions, and suggest directions for future work.

Chapter 2

Computational Approaches for Visual Object Recognition

The computational approach to human vision goes back to the nineteenth century when the algebraic formulas for predicting perceived hues from the spectral energy distributions, perceived sizes and shapes from retinal images, perceived depth from image disparities between the left and right eyes, and perceived brightness from simple luminance distributions were formed [13]. Although we still do not have a model of recognition powerful enough to come close to matching the capabilities of a human, many plausible theories and models of visual object recognition have been proposed. The theories and computational models of visual object recognition described later in this chapter have different explanations of high-level processing – how knowledge about objects and the world is stored internally, how the information extracted from the sensory input is represented, how memory can be activated under varying conditions, and how the representation of the input is matched against representations of objects in memory. The two dominant paradigms are the *model-based* or *view-invariant* approach and the *appearance-based* approach.

2.1 Model-Based Approaches

In model-based approaches, objects are represented as 3D geometric models, and pose constraints direct the process of matching abstract image features to model features [115, 20, 92, 93, 66, 19]. Therefore, the observer's viewpoint is assumed not to affect his perception of the object. This approach goes back to one of the most influential books on object recognition, David Marr's *Vision* [96]. According to Marr, objects are recognized by matching salient 3D features of a scene to abstract models containing lists of these features and their interrelations. The processing is accomplished through a sequence of stages: the *primal sketch* which contains significant changes in luminosity across the image, a $2\frac{1}{2}D$ *sketch* which specifies for each portion of the visual field the depth of the corresponding distal object and the local orientation of the surface at that point, and, finally, the full 3D representation of space and objects within it.

Other than Marr, there are various researchers employing the model-based approaches to object recognition. Brooks introduced a vision system called *ACRONYM* [28], representing the first significant effort to build full 3D model-based system based on the parameterized representation of an object. Grimson intensely studied the role of geometric measurements and constraints in determining the pose of object in the scene and the correspondence between image features and model features [56]. Finding optimal correspondence and a pose of a 3D object is the core part of model-based approaches. Beveridge & Riseman [19] proposed search algorithms that efficiently solve the problems under full 3D perspective.

Among the notable successful computational work based on full 3D models of objects are Lowe [92, 93] and Huttenlocher & Ullman [66]. Lowe's model is directed primarily toward determining the orientation and location of objects, even when they are partially occluded by other objects, under conditions in which exact 3D object models are available. For a viewed image, edges are detected by finding sharp changes in image intensity values across a number of scales, and then grouped according to viewpoint-invariant properties – collinearity, parallelism, and proximity. A few of these image features (edges) are matched against those of the object model generated from a particular orientation of the object that would maximize the fit of those image features. Then the location of additional image features are proposed and their presence in the image is evaluated.

Whereas Lowe's model is limited to images with straight edges, Huttenlocher & Ullman's model has the potential for recognizing a broader class of objects, including those with curved surfaces[1]. It has somewhat similar characteristics to Lowe's model. All the object models that are candidates for possible matches for the image are aligned (rotated) before they are matched with the image and tested for geometric fit. This alignment model offers a possible explanation for those cases in which recognition depends on re-orienting a mental model.

In the field of psychology, Biederman introduced a theory of human visual object recognition called *Recognition by Components* (RBC) [20]. Instead of using full 3D models of objects, RBC models objects as combinations of volumetric primitives called *geons* and matches the primitives and their interrelationships extracted from images to those of object models to recognize an object. To determine a set of geons present in the scene, Biederman adopts the non-accidental instances of viewpoint-invariant properties, such as collinearity, curvilinearity, symmetry, parallel curves, and co-termination, introduced by Lowe. Since non-accidental properties are generally viewpoint invariant, geons can be differentiated by their invariant properties in the 2D image. The use of those properties for generating and representing geons is supported by theoretical and empirical evidence, as well as psychological ev-

[1]Later, Lowe extended his model so that it can be applied to images including curved surfaces as well [94].

idence [20]. Biederman also showed that viewpoint-invariant properties are employed by humans to achieve invariance in their recognition of novel objects at new orientations in depth [22]. Once the arrangement of geons is extracted from the image, it is matched against that of objects in memory. The simplicity of geons with the largely viewpoint-invariant properties makes the recognition relatively robust when the objects are rotated in depth, novel, or extensively degraded [21].

An example of recognition-by-parts had also been proposed by Pentland [115] in the computer vision community, who used deformable implicit functions (superquadrics) to model objects. Biederman's RBC theory was adapted by the object recognition community and several geon-based vision systems have been introduced. Among them are Bergevin & Levin's PARVO (Primal Access Recognition of Visual Objects) [37] and OPTICA by Dickinson *et al.* [38]. Also, Biederman proposed his own implementation of geon theory called JIM, using a neural-net model [65]. Although these systems show some practical use of geon theory, Dickinson mentioned that there remain obstacles for realizing successful geon-based recognition: the recovery of geons from real imagery, the difficulty in explicitly modeling real objects using geons, and the lack of representational power provided by geons for the task of interacting with the world [37].

Whether representing objects using complete 3D models or a structural description specifying the relationships among viewpoint-invariant volumetric primitives, model-based approaches have a number of problems. First, the modeling systems put limitations on the types of objects that can be recognized, and second, acquiring accurate 3D models of objects is often a very difficult task. In most cases, model-based approaches use human-made models and/or require CAD-like representations, but such representations are not always available, especially for non-rigid objects. Other problems are unreliable feature extraction methods and the combinatorics of feature matching for constructing 3D shapes from images.

In addition to the computational problems, there are psychological arguments against model-based approaches. Pizlo [117] dismisses model-based approaches on the grounds that they rely on depth cues which he claims are unimportant, since in their absence we can still recognize shapes. This oversimplifies matters somewhat because model-based schemes such as Biederman's [20, 22] can also be applied to objects without textural or other depth cues, and because the human visual system may be redundant; removing one source of information may not necessarily imply sudden failure. However, empirical evidence of a monotonic relationship between recognition performance and viewing angle provides further evidence against model-based schemes.

The correlation between recognition time and the object's disparity from a previously learned pose was first reported by Shepherd & Metzler [135]. It can be interpreted as evidence for mental rotation of internal 3D models of objects; however, it has been shown that the recognition accuracy drops as a function of orientation disparity from a learned view [127], which contrasts to the predictions of model-based theories. Similar results were reported by Bülthoff & Edelman [29], and they also showed that if the orientation of an object falls between the two previously learned views, it can be recognized better than when it is outside of the two views. From this theory, they proposed an object recognition model that uses nonlinear interpolation of stored 2D views [43].

Recently, similar psychophysical results have found in more thorough experiments by Tarr and his colleagues [144]. They found a pattern of viewpoint dependence which systematically related to the distance from the previously trained views for both 2D and 3D object recognition rotated in the image plane and in depth [143, 146]. Also, it is shown that the viewpoint dependent mechanisms are involved in both basic- and subordinate-level recognition [50, 60, 61]. They conclude that viewpoint dependent processes can be generalized for a range of recognition tasks with different levels of recognition goals.

2.2 Appearance-Based Approaches

Appearance-based approaches model 3D objects as a set of 2D images where each of the images corresponds to a specific view of the object. Therefore, they dispense with the need for storing explicit 3D models and recognize objects by matching the input image against the stored views in the set. In other words, appearance-based approaches consider object recognition to be an image retrieval problem, while model-based approaches view it as geometric-model retrieval problem. Since appearance-based approaches may make the recognition process faster, more general and robust, and also make it easier to obtain training data, the interest in appearance-based techniques has grown quickly. As a result, many appearance-based theories and methods have been proposed. They can be categorized roughly into three methods: view interpolation, feature space matching, and subspace projection methods.

2.2.1 View Interpolation Theory

In view-interpolation theory, recognition is generalized to novel views by linear/non-linear interpolation of training views. As described in the previous Section, Edelman & Bülthoff's work [43] using non-linear interpolation of stored 2D views falls into this category. Other notable models are those of Poggio & Edelman [119] and Murase & Nayar [108]. Poggio & Edelman described a view-interpolation theory of recognition that is particularly well-suited to the constraints imposed by biological implementations. Their model is based on the mathematical observation, described by Ullman [153], that the views of a rigid object undergoing transformation such as rotation in depth reside in a smooth low-dimensional manifold embedded in the space of fixed 2D views of the same object. When the stimulus view of an object is presented intermediate receptive field responses – measurement-space distance between the stimulus view and the stored views – are formed using Gaussian radial basis functions (RBFs) centered at the stored views. Then, the responses are used to linearly interpolate stored views. If enough stored views are available, the model can account for the variability in pose of the target object.

Murase & Nayar's model is similar to Poggio & Edelman's approach, except that the low-dimensional manifold is formed using principal components of an image training set. The connection between principal components and the low-dimensional subspace called *eigenspace* associated with the training images is described in Section 2.2.3. In Murase & Nayar's approach, two types of subspaces are used: the universal eigenspace formed with all images in the learning set, and the object eigenspace computed from individual object image sets. The appearance representations of objects in an eigenspace describe a smoothly varying manifold. Murase & Nayar use a standard cubic-spline interpolation algorithm to compute the manifolds in both universal and object eigenspaces. An input object is recognized by finding the closest manifold in the universal eigenspace. Once the object's identity is known, it is projected onto the corresponding object eigenspace to estimate the pose by computing the parameters that minimize the distance between the projected point and the manifold.

2.2.2 Feature Space Matching Methods

In this approach, objects are represented by feature vectors, and recognition is achieved by matching the feature vector computed from an image with stored model features. This is by no means a new approach. It has been widely used in traditional pattern recognition, where the goal is to find decision boundaries in the feature space that separate patterns belonging to different classes. Also, it shares a common mechanism with model-based approaches in that features are extracted from input and compared with stored model features. However, unlike the model-based approaches, the stored features in appearance-based approaches are all extracted from 2D views, therefore no 3D model extraction from input features is involved. The main concerns in the feature-space matching method are what kind of features are salient (i.e. discriminant), how to combine the different types of features, and how to match them with the stored feature vectors.

Rao & Ballard proposed an active vision architecture in which an image is represented as a high dimensional vector of responses to an ensemble of Gaussian derivative spatial filters at different orientations and scales for fast computation of visual routines [126]. To identify an object, image and model response vectors are compared using a similarity metric called normalized dot-product (or correlation) and a straightforward voting process is used to determine the winning model. The small changes in viewing position which causes changes in a few individual filter responses are ameliorated by the reliance on a large number of filter responses. As in most other appearance-based methods, significant changes in viewing angle are handled by storing feature vectors from multiple views.

Mel represents a view of an object with a set of feature channels in his 3D object recognition system called SEEMORE [100]. The features used in the system are those that are sensitive to object identity, such as an object's color, shape, or texture, and relatively insensitive to changes unrelated to object identity, such as pose. Each feature channel is the sum of responses of elemental nonlinear filters, which are parameterized by position and internal degrees of freedom, over the entire image. The training views cover multiple viewing angles and scales for each object. A nearest-neighbor classifier finds the closest match between the observed feature vector and the stored models.

Another feature-based approach to be noted is Schmid & Mohr's method of combining greyvalue invariants with local constraints [132, 133]. In this method, features are computed by applying differential greyvalue invariants [80] at several scales to interest points. These features locally characterize the input and, since the interest points are the locations with high information content, they are highly discriminative. Schmid & Mohr use a voting scheme and multi-dimensional hash table for robust and fast matching. To reduce false matches, they also add a simple constraint that specifies the geometric relationship between neighbor interest points. Rotation in depth is handled by storing multiple views for each object.

2.2.3 Subspace Projection Methods

In subspace projection methods, unknown images are projected onto a space formed by basis components of an image data set and similarity is measured between the projected representations. The most popular procedure used is *Principal Component Analysis* (PCA). PCA builds a global linear model of the data set, which is an n dimensional hyperplane spanned by the leading n eigenvectors of the covariance matrix of the data set. The number of eigenvectors, n, is determined by the amount of error that can be tolerated. PCA produces an optimal linear basis in that the expected squared distance between an input and its reconstruction from an n dimensional encoding is minimized. Since n is generally smaller than the dimension of image space, PCA has been commonly used for compression and encoding as well as for object recognition. Kirby & Sirovich [79] first showed that PCA is an optimal compression scheme for a set of images and Turk & Pentland [151] were the first to apply PCA to face or object recognition. Later, Murase & Nayar [108] applied PCA for learning complete parameterized models of objects. As described earlier, a set of images of an object are projected onto eigenspace and a manifold, which is parameterized by pose and illumination, is formed by interpolating projected views. Their method has been successfully applied to recognize more general objects with complex appearance characteristics [110].

Factor Analysis (FA) is a statistical technique similar to PCA for explaining the variance in a data set in terms underlying linear factors. FA was originally developed in social sciences and psychology, where the major use of FA is to develop objective tests for measurement of qualities such as personality and intelligence [139]. Its goal is to explain the correlations among a set of observed variables in terms of a smaller number of relevant and meaningful factors. A single global FA model, however, has not been widely exploited for object recognition. Instead, recent work on recognition tasks fits mixtures of factor analyzers to data sets using EM algorithm [47, 55, 63].

Linear Discriminant Analysis (LDA) has also been used for computing the basis vectors for a data set. Given the class assignments of objects in the training data set, LDA finds the discriminant axes that maximize between-class scatter while minimizing within-class scatter. When the number of classes is c, such axes are defined by the $c-1$ eigenvectors associated with the largest eigenvalues of a matrix formed by multiplying the inverse of the between-class scatter matrix and the within-class scatter matrix. Therefore, the problem becomes mathematically the eigenreduction of a real-valued matrix as in PCA. LDA has been used for finding discriminant features for image retrieval [137, 138] and face recognition [161].

Recently, another procedure called *Independent Component Analysis* (ICA) [34] has been used for face recognition [9, 122]. While PCA decorrelates the signals, ICA performs a linear transform to make the resulting variables as statistically independent from each other as possible. Therefore, the basis axes in ICA are not necessarily orthogonal to each other. ICA first received attention in signal processing, where it has been used to recover independent sources given sensor observations that are unknown linear mixtures of unobserved independent source signals [34, 15]. Later, Bell & Sejnowski [16] proposed that the independent components of natural scenes are localized and oriented edge filters similar to Gabor filters. More recently, ICA has been applied for representing high dimensional data for object recognition and classification [27], and comparative studies have been performed between ICA and PCA for face recognition [6, 9, 7, 10, 90, 105, 159] and facial expression coding [7, 8, 39].

2.2.4 Other Appearance-Based Methods

Although appearance-based object recognition methods have recently demonstrated good performance on a variety of problems, they also have some restrictions. Many methods that use basic image features to hypothesize the identity and pose of objects in a scene need to compute correspondences between image features and model features. The complexity of determining feature correspondence grows exponentially with the number of extracted image features – this is the same case as in the model-based approaches. Moreover, the image feature extraction and grouping processes are unstable, often producing broken and spurious features. Also, many appearance-based approaches require good figure-ground segmentation of the object, which severely limits their performance in the presence of clutter, partial occlusion, or background changes.

More recently, other appearance-based approaches have been proposed to overcome many of these problems. Among them are Schiele & Crowley's *multi-dimensional receptive field histogram matching* [131] and Nelson's theory of using 2-stage associate memory for recognizing 3D objects [111]. Although both approaches use local features, recognition is not achieved simply by matching corresponding features. Shiele & Crowley's approach is motivated by the color histogram work of Swain & Ballard [136], in which objects are modeled by their color statistics. Shiele & Crowley represent objects using joint statistics of local characteristics. The probability density functions for local characteristics are approximated by multi-dimensional histograms and recognition is achieved by comparing probability distributions using histogram matching or by computing probabilities for the presence of objects based on a small number of measured local characteristics.

Nelson's approach combines an associative memory with an evidence combination technique. The basic idea is to use distinctive local features called '*keys*' and two stages of a general purpose associative memory. The recognition system uses key features to extract hypothesis for the identity and configuration of all objects in memory that could have produced such features. The second stage associative memory takes the hypotheses and groups them into clusters that are mutually consistent within a global context. This step is keyed by configurations that represent 2D rigid transforms of specific views. The system lists object identity and pose hypotheses. Since the system uses merged percepts of local features rather than the complete object appearance, it is less sensitive to background clutter and occlusion.

2.3 Summary

Over the last decade, there has been tremendous progress in visual object recognition. Researchers in various fields of study have proposed a large number of theories and models. Those described in this chapter are just a part of the larger literature. However, they show the prominent works in the two dominant, long debated paradigms for visual object recognition. Although we are still short of a general model, a body of work in psychology and psychophysics [29, 43, 127, 135] provides converging evidence for view-based representations of objects in the human visual system, and therefore support appearance-based approaches as a more plausible candidate and more relevant to biological systems.

Chapter 3

The Ventral Visual Pathway and Expert Object Recognition

In this chapter, we provide a detailed description of Kosslyn's model of the ventral visual pathway. The functional role of each component in the ventral visual pathway is discussed with the biological evidence supporting it. We also review studies on the existence of an expert object recognition pathway mostly described by Tarr and his colleagues.

3.1 Kosslyn's Functional Model of Ventral Visual Pathway

Kosslyn's model of object identification summarized in the first chapter involves a broad range of research areas, covering almost every aspect of a vision system. For example, it includes integration of 2D and 3D processing and knowledge-base maintenance. In this study, we focus on visual object recognition without 3D modeling. In fact, this can be considered as a computational goal of the ventral system in Kosslyn's model. In this chapter, we provide a more detailed description of Kosslyn's model of the ventral visual pathway shown in Figure 3.1.

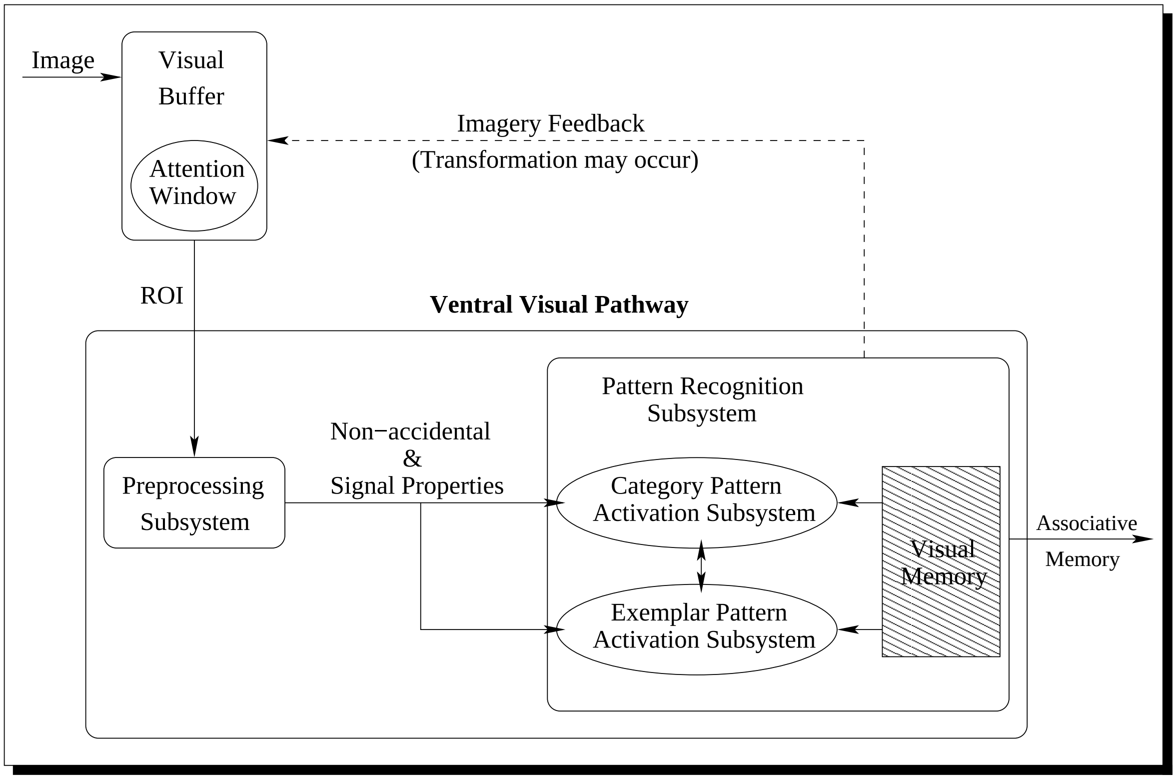


Figure 3.1: Initial processing components and ventral visual pathway of Kosslyn's model.

3.1.1 Visual Buffer

The visual buffer refers to the first cortical area that receives signal from the eyes. (This area is known as area V1, primary visual cortex, striate cortex, OC, and area 17 ([82], pp. 13). See the glossary in Appendix C.) In Kosslyn's model, the visual buffer is the place where images are produced as a pattern of neuronal activation. It can be activated not only through the signal from the eyes, but also by information stored in memory. The spatial structure of the visual buffer is "retinotopic" in the sense that the neurons in this area are organized to preserve the arrangement of a visual image on the retina – nearby neurons are activated by the stimuli in adjacent positions in the visual field.

Along with PET study results [46], striking evidence of the retinotopic structure of the area is provided by Tootell *et al.* [150]. They injected a Macaque monkey with 2-deoxyglucose (2DG) and flashed a visual stimulus composed of eight rays and five logarithmically equally spaced concentric rings for 25 to 30 minutes. The animal was then sacrificed and the surface of the striate cortex was smoothed out. The 2DG map of the surface showed that the pattern of brain activation in the cortex preserved the structure of rings and rays in the input stimulus (see [150] or pp.14 in [82] for the resulting 2DG map). Hence, the content of the visual buffer is considered to be represented as transformed versions of the retinal image.

The sophistication of the image-like representation in the visual buffer, however, should not be underestimated: it appears to be a multi-scale pyramid. Adelson *et al.* [2] proposed an image representation based on certain aspects of early information processing, such as sensitivity to spatial and size frequencies, in the human visual system. They proposed images be represented as the responses to basis functions, which resemble the receptive fields in the human visual system, of many sizes and locations [2]. Later, Burt & Adelson presented an efficient encoding scheme of an image as a Laplacian pyramid [30]. Marr [96] also considered a multi-scale representation for the edge detection process in his theory of vision. It is also known that multi-scale representations are useful for many low-level computations [1].

3.1.2 Attention Window

More information is present in the visual buffer than can be processed in detail, and an object can appear at different positions, sizes, and orientations on the retina. Thus, a neural mechanism is required for directing our attention to a useful portion of the visual field to accomplish the task at hand. The role of the attention window in the model is to select a region in the visual buffer. The pattern of activation within the region is passed through deeper into the system for detailed processing. The attention window can be shifted rapidly to any part of the visual buffer, and it can also be scaled to different size ranges [17, 82].

Psychological evidence for an internal adjustment of the attention window comes from the cueing effect. Eriksen and Hoffman [45] show that, if a target location has been previously cued, then the response time for detecting the target is faster than when the location is not cued. In their experiment, the cueing interval was shorter than the time needed for eye movement, thus providing a evidence for covert attention shifting.

Another experiment performed by Cave and Kosslyn [32] showed the subject diamond and rectangular shaped stimuli successively. The subject was asked to decide whether the sides of each stimulus were of equal length. The resulting response time increased linearly with increasing size disparity. That is, when the second stimulus appeared at an unexpected size, the subject had to adjust the size of the attention window to cover the region occupied by the stimulus. In fact, Kosslyn hypothesizes that the size of the attention window is adjusted by selecting the appropriate level of scale of the multi-scale representation in the visual buffer ([82], pp.95).

Anderson and Van Essen presented a biologically plausible model of attention shifting [4]. They introduced "shifter circuits", which control information flow in the visual pathway by dynamically linking and aligning arrays of neurons in different levels. The circuits also preserve local spatial relationships. The shifter circuit successively maps focused regions to the cortical module in higher levels until it reaches a single "attention center" at the highest level. In this way, information from the attention window located at an arbitrary position in the visual buffer is routed to a high-level cortical area. Later, Olshausen *et al.* added a scaling mechanism and autonomous control for shift and scale [113].

3.1.3 Preprocessing Subsystem

The pattern of the region within the attention window is sent to the preprocessing subsystem, which extracts object properties, such as shape, color, and texture. These properties are then sent to the pattern activation subsystem where matching to the stored representation occurs.

As mentioned earlier, information received in the visual buffer produces a pattern of activation. Thus, it is possible that different objects are distinguished by the differential activity of this area. However, the visual buffer is neither homogeneous nor isotropic, which means that identical objects at different positions, sizes, and orientations activate different cells in the area, projecting different input images onto the visual buffer. Therefore, object properties that are relatively invariant under scale, translation, and rotation are required. These properties have been called *non-accidental properties* [91], since it is highly unlikely that an accidental alignment of eye and object features would produce such properties.

Non-accidental properties include collinearity of points or lines, curvilinearity of arcs, shape symmetry, parallel curves, and co-termination [20]. The use of these properties in visual recognition is well supported by theoretical and psychological evidence. The principle of perceptual organization is that certain properties of edges in 2D images are taken by the visual system as strong evidence for the same properties in the 3D world [91]. Results from a naming experiment performed by Biederman also provide supporting evidence [20]. In the experiment, subjects were shown drawings of objects with part of the contour removed; in one case all the non-accidental properties were intact, in the other case, the non-accidental properties were deleted. The subjects spent more time naming the object when the non-accidental properties were removed.

Since non-accidental properties describe what is invariant under viewpoint changes, they are powerful and relatively general. Unfortunately, some objects may have identical non-accidental properties despite being different. (An example given by Kosslyn is a pen and a mechanical pencil [82]). When one needs to distinguish a specific object from those having the same non-accidental properties, other information should be used. Also, there are many non-rigid natural objects (e.g. trees and humans) that can not be easily described by non-accidental properties. In circumstances, where non-accidental properties are non-discriminating, other characteristics have to be used for recognition. Kosslyn refers to these characteristics as *signal properties* [82]. The preprocessing subsystem extracts signal properties via two different principles. First, perceptual units, such as regions of homogeneous color and texture and sets of contiguous elements, are extracted. While the perceptual units are organized by bottom-up processing in the visual buffer, the preprocessing subsystem can also be tuned through top-down processing to extract properties specific to the object to be recognized.

An interesting example that shows the use of signal properties in a human visual system is the empirical experiment on sexing of day-old chicks performed by Biederman and Shiffrar [24]. The subjects were shown pictures of cloacal regions of male and female chicks, and asked to judge the sex of each chick. After they were trained to attend to the shape of a critical cloacal structure, the accuracy was increased compared to the pre-training case. Moreover, their analysis more closely resembled that of the professional sexers after training. This suggests that object specific characteristics can be learned and used for perceptual distinctions.

3.1.4 Pattern Activation Subsystem

The input is finally matched with the stored information in the pattern activation subsystem. There are two important issues regarding this process: the type of representation to be used to store information, and how the input is matched to the stored representation.

The pattern activation system stores information in a distinct physiological structure called *visual memory.* It has been shown in monkeys that the infero-temporal area of the cortex, which appears to store visual memories, is organized as repeated sets of columns [48]. Fujita *et al.* recorded cell responses to sets of stimuli through the vertical penetration in the cortex as well as across the surface. The results show that neurons with similar selectivities span most of the cortical layers and are clustered in patches across the cortical surface. Moreover, they found that the preferred stimuli shared by vertical columns of neurons were not restricted to shapes, but could be other visual features of objects such as gradation [48].

Kosslyn argues that the representation in visual memory should include a wide range of properties so that the image can be reconstructed from the representation. Non-accidental and signal properties are used to index these representations. From the columnar structure in the infero-temporal area of cortex, Kosslyn suggests that the long-term representation of an image would be a feature vector across the functional columns, called a *compressed image* [82]. According to Kosslyn, the compressed image representation lacks topography; however, his intention may be the lack of explicit physical topographic structure, since topographic information has to be implicitly included in the representation to be able to reconstruct the image. The precise representation of compressed images has not been defined; however, it seems that the representation should be determined according to the reconstruction scheme. For example, if coarse coding is used to reconstruct the image, the relative strengths of inputs from larger regions have to be stored in high-level visual areas [82].

The pattern activation subsystem is divided into two separate, more specialized subsystems called the *'category pattern activation subsystem'* and the *'exemplar pattern activation subsystem'*. The category pattern activation subsystem encodes category information and classifies an input as a member of a visual category, while the exemplar pattern activation subsystem stores information about instances, and therefore recognizes specific exemplars. This hypothesis is based on the idea that characteristics that are critical for identifying particular instances may not help to classify them as a member of a category, so the system must put aside that information when it categorizes the input. In other words, the information-category and information-exemplar mappings may be incompatible.

There have been many experiments conducted to show that exemplar recognition and categorization are performed in separate parts of the brain. For example, Marsolek [98] conducted a visual pattern categorization task and found that the performance for categorizing previously unseen prototypes is better when they are shown initially to the left hemisphere of the brain. Similar results were reported by Vitkovitch & Underwood on object/non-object (combination of object's parts) determination tasks [155] and by Sergent *et al.* on a PET study [134]. Milner's study of patients whose right temporal-lobes were removed also reported that patients showed a selective deficit in memory with the tasks involving specific exemplar matching [103]. This empirical evidence suggests that the right hemisphere plays a special role in the representation of specific exemplars, while the left hemisphere is more involved in accessing stored categorical representations.

Jacobs & Kosslyn further hypothesize that higher resolution encodings are needed for exemplar identification than for categorization; therefore, relatively large overlapping receptive fields are used for recognizing specific exemplars [72, 82]. They verify their hypothesis by showing that the receptive fields of artificial neural networks learned to encode specific exemplars are larger and overlap more than the receptive fields of networks trained to categorize shapes [72]. Kosslyn listed other experimental evidence that supports the difference of receptive field sizes according to the role of the subsystem ([82], pp. 182–185). He also notes that size differences do not always alter performance; rather, it depends on circumstance and the nature of the stimuli. For example, simple detection tasks show no difference in performance between the left and right hemispheres. So the hemispheric differences may reflect attentional biases rather than the number of hard-wired connections to neurons with different-sized receptive fields.

3.1.5 Imagery Feedback

The properties extracted from the input by the preprocessing subsystem and their relative positions are matched to those of stored visual representations in the pattern activation subsystem. If the properties are distinctive enough, the representation of the best-matching object is strongly activated, and recognition occurs. However, if the activation is not strong enough, the best-matching object sends feedback to the visual buffer. The feedback generates a mental image in the visual buffer [85], which augments or "fills in" the input image. Indeed, the mapping from the representation to the visual buffer is continuously adjusted until the feedback augments the input as well as possible ([82], pp.145).

This top-down, imagery feedback process is similar to Lowe's idea - generate an image by activating a stored model and compare it to the input image [91, 92, 93]. Why not simply match the input bottom-up to the stored image representation, when the extracted features are too weak to make a strong prediction? According to Kosslyn, matching compressed images (the internal representation) directly is the same as simple template matching, and therefore has all the problems that approach entails (especially with the variations of object views). He also considers deformable templates, *à la* Ullman [152], as being incompatible with the anatomical structure of the brain; there is no topographic structure in the cortical area that stores visual memories [36]. Instead, he suggests that the encoded compressed image representation is used to reconstruct the image so that information that is implicit in the stored representation can be accessed. For example, if local geometric properties of a shape are implicit in the compressed image representation, it would be difficult to generalize recognition over variations in the properties when the comparison is made with the stored representation. The reconstructed image in the visual buffer makes those properties explicit, therefore makes recognition of the object more general [82].

Several experimental studies involving imagery feedback in visual perception have been conducted. Cave and Kosslyn [32] use simple superimposed geometric stimuli to show that stored shapes are adjusted to match the input. Koriat *et al.* [81] asked subjects to identify Hebrew letters with orientation changes. In their experiment, the response time increased with greater orientation disparity from the preceding letter, and this only happened when the previously seen letter is the same as the letter currently showing. Furthermore, the time with large orientation disparity was still faster than the time for identifying different letters, which suggests that the previous letter may cause a priming effect and may be used for imagery feedback. In other words, less bottom-up processing makes it faster. Michelon and Koenig [101] also tested the participation of imagery feedback in perception using a priming paradigm.

Besner conducted an experiment [18] suggesting that imagery feedback is used in some cases, but not all the time. In their experiment, subjects were asked to judge whether a pair of stimuli were the same or different. In the first case, the judgment had to be made between a shape and the same shape with different orientation and, in the second case, the discrimination was between different shapes. That is, for the first case, “different” was required when the stimuli had the same shape but a different orientation while different shapes made “different” classification for the second case. In the first experiment, the response time increases with the size disparity equally on the “same” and “different” trials. In contrast, in the second experiment the size disparity had less effect on the “different” trials than “same” trials. Kosslyn interprets the results as follows: If the shape is different, the non-accidental properties are distinctive enough to make a decision based on the initial matching process, therefore no imagery feedback occurs. If the shape is the same, then the initial matching process can not distinguish the stimuli based only on non-accidental properties, therefore imagery feedback is generated. In this case, it takes more time because the mapping is adjusted until the size and orientation disparity are matched.

It should be noted that the description of the pattern activation subsystem in the overall recognition system is imprecise. As discussed above, it is not clear what kind of representation is used for matching properties, or how the matching process is accomplished. Kosslyn doubts that a template-like image is fully generated and compared to the input. He says, "rather, a better way to conceive of this imagery-feedback process involves the concept of vector completion" ([82], pp.121). Kosslyn also says that vector completion typically does not produce a fully formed image, therefore differs from template matching; however, he never describes what it actually measures.

3.2 Expert Object Recognition

In [102], Milner and Goodale have argued for modularity within the dorsal pathway: anatomically distinct regions in the dorsal pathway are modularly arranged according to their roles in achieving different components of actions. By implication, the ventral stream should also be composed of multiple, anatomically distinct components performing different functional roles. This hypothesis is verified by brain imaging studies, which show preferential activity patterns to different classes of objects in discrete areas of the ventral pathway. Ishai *et al.* used fMRI to compare activation in the ventral occipital and temporal cortex of normal subjects while they observe images of faces, houses, and chairs. The results showed different regions responded preferentially to images of objects in different categories [70, 71]. Similar results were found using PET scan while performing face and scene processing tasks [109], and using fMRI during mental imagery of faces and places [112].

Recently, brain imaging studies for functional specialization in the human cerebral cortex have concentrated on the face recognition task. In many studies, a region in the lateral portion of the fusiform gyrus is consistently reported as a face-specific area which responds more to faces than to other objects. For example, fMRI images scanned while patients were shown images of human faces revealed activation in the fusiform gyrus in addition to the primary visual cortex [33, 77, 125]. Subsequent PET studies confirmed the activation in the fusiform gyrus, while adding another locus of activity in the right inferior frontal gyrus, an area previously associated through lesion studies with visual memory [95]. Moreover, in both the fMRI and PET studies, the activation was unique to the task of face recognition. Images of places triggered another, distinct pathway with activation in the parahippocampal gyrus [109, 112].

The specialization for processing faces is also observed in psychological and behavioral studies. For example, Tanaka and Farah reported that detection of the difference in individual face parts is facilitated when the entire face is presented. This holistic or configural effect was not found for non-face objects or inverted faces [141]. Kalocsai and Biederman also showed the configural characteristics of face recognition by comparing the recognition performance for complementary pairs of images of faces and non-face objects created by having every other Fourier component [74]. These findings in neuroimaging and behavioral studies led to speculation that evolution had created a special visual pathway for recognizing faces, and the locus of activation within the fusiform gyrus was dubbed the *fusiform face area* (FFA) [78].

The view of face-specific mechanisms, however, has been challenged by more recent studies. The argument is that the evaluation for face-specific mechanism is incomplete in that those previous studies did not consider all the factors that play a role in determining visual recognition behavior [142]. For example, in many cases, the evaluation was conducted for different level of perceptual categorization – non-face objects recognition often involves categorical-level judgments while face recognition involves individual-level judgments – and the degree of perceptual expertise with a given class was different: humans are highly experts in recognizing faces compared to other general objects.

Of particular interest in the argument is the effect of domain expertise. An electrophysiological study employing event-related potential (ERP) showed that the early negative component N170, whose magnitude is significantly larger during face processing, was found when dog and bird experts performed categorization in different levels for objects in their domain of expertise [140]. In fact, N170 was larger when the task was performed in the domain of expertise than outside of the domain of expertise. These results indicate that face-specific mechanisms are also recruited for expert object recognition.

Most notably, a body of work on perceptual expertise by Tarr and Gauthier, and their colleagues provides very thorough and strong evidence for a mechanism specialized for recognizing objects in the expert domain. They factored in the past experience of their subjects, and found FFA activation in dog show judges when they view dogs, and in bird experts when they view birds [145]. Most convincing of all, Tarr and his colleagues show that as people learn to become expert at recognizing a class of objects, their recognition mechanism changes [49, 51, 53, 145, 157]. They created a class of characters called *greebles*, which in addition to individual identities can be grouped according to gender and family. When novice subjects view greebles, fMRI results show that their FFAs are not active. The subjects are then trained to be experts at recognizing greebles, where the definition of expert is that they can identify a greeble's identity, gender, or family with an equal response time. After training, the FFAs of these subjects are active when they view greebles. It is therefore reasonable to conclude that the previously known face-specific pathway is recruited more generally for expert object recognition, and the FFA is part of the pathway.

To model and evaluate expert object recognition, it has to be specified what properties constitutes expert object recognition. Gauthier and Tarr use categorizing objects at multiple levels with equal time as a defining characteristic of expert recognition in their greeble studies [54]. For example, people recognize faces they have never seen before as being human faces and, at the same time, they can instantly identify familiar faces. Another property of expert object recognition is that it is viewpoint dependent. The response of the FFA to images of faces presented upside-down is minimal [59]. The FFA responds to faces viewed head-on or in profile, but not to images of the back of the head [149]. In [52], upright and inverted greebles are presented to both novices and expert subjects. While novices do not show activation in FFA, the activation pattern is remarkably different for the greeble experts and clearly shows the presence of FFA in fusiform gyrus.

The second property of expert object recognition indicates that recognizing familiar objects from familiar viewpoints is different from recognizing objects from novel or unusual viewpoints. Since our work focuses on modeling the expert object recognition pathway in the ventral visual stream, we assume that the system can be trained for given classes of objects viewed from fixed viewpoints as humans learn from experience. The system's structure is designed following Kosslyn's ventral stream model, except for the attention window and imagery feedback mechanisms. Expert object recognition also indicates that the same pathway may be recruited by different classes of objects as people gain expertise for objects, even though the visual features are very different. This justifies our application of the same underlying mechanism to recognizing objects in different domains in the experiments performed in this study.

Chapter 4

Early Stages of the System

In the preceding chapters, we described computational approaches and biological models of object recognition. In this chapter and the next, we describe an algorithmic implementation of Kosslyn's model of visual object recognition in the context of expert object recognition. The computational methods used for implementing the proposed system are described in relation with the biological theory of vision for the corresponding functional role.

Kosslyn's model of visual object recognition described in Section 3.1 is divided into three components. An initial processing component provides inputs to the ventral visual pathway. At the beginning of the ventral pathway, a preprocessing subsystem extracts features and passes them to the pattern recognition subsystem, which performs classification and recognition of the input objects (Figure 3.1). In this chapter, we first describe the initial processing as image patch extraction and transformations. Then, we describe the features used and how those features are generated from the data provided by the initial processing component (Figure 1.4). The next chapter will describe algorithms for classification and exemplar recognition.

4.1 Patch Extraction

As shown in Figure 3.1, the input to the ventral visual pathway is the visual buffer, including but not limited to the attention window. In perception, the visual buffer receives far more information than can be processed in detail. Also, an object in a scene can appear at different locations, sizes, and orientations on the retina. The attention window is capable of selecting a region in the visual buffer and can be scaled to different size ranges. The information in the selected region with specific scale is further processed as it pass through the cells in the visual buffer so that the end product of early vision is generated, which are the scaled and filtered versions of the retinal image centered on a fixation point. In this study, we do not directly model the attention window because we do not know how to model its operation algorithmically, and because attention is a super-visual task involving non-visual inputs (see Section 3.1.2). Instead, we assume that the attention window produces a set of unlabeled patch images.

To extract patches from an input image, the system exhaustively searches the input by running a corner detector[1] across the image. Then, it focuses sequentially on every point, generating image patches that are rotated according to the dominant edge orientations[2]. Figure 4.1 shows the detected interest points and the extracted patches from two locations with different orientations.

The role of patch extraction is consistent with that of a biological attention window in that it selects regions that will be sent down the ventral pathway for further processing. It should be noted, however, that the mechanism does not follow any biological model of attention. Computational models of attention have been proposed by other authors. Posner and colleagues proposed a spotlight-like operation,

[1]For a block of pixels, a matrix $[\sum I_x^2, \sum I_x I_y \ ; \sum I_x I_y, \sum I_y^2]$ is formed where I_x and I_y are the first derivatives with regard to the x and y axis within the block. Then, the eigenvalues of the matrix is computed. Interest points found are corners with large eigenvalues in the image [58].

[2]Each image in all data sets, except aerial image data set, used in this study contains one objects and approximately registered. Therefore, corner detector runs only on aerial images, so the scale is fixed.

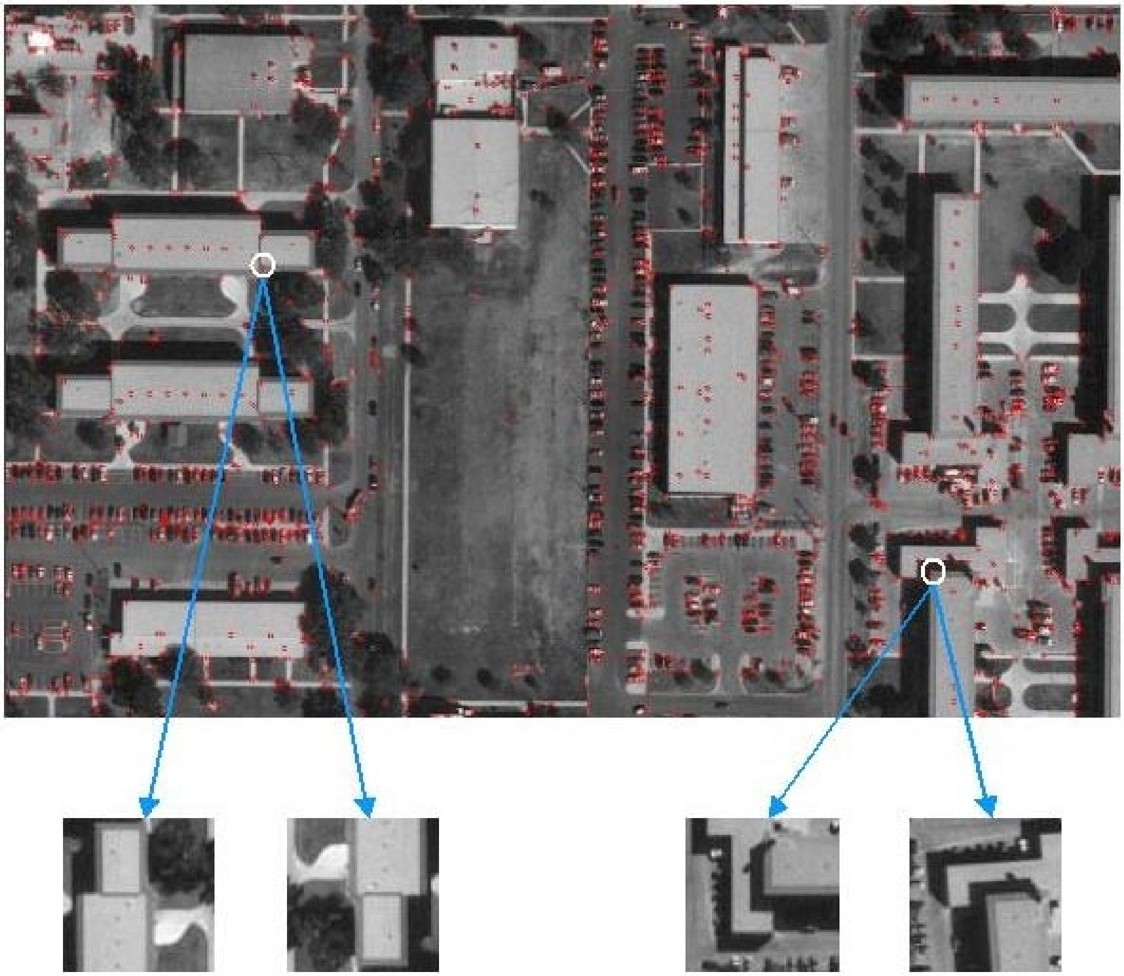


Figure 4.1: Patch extraction from an input aerial image. The interest points are shown with red circles on top of the image. Sample patches extracted from two locations and rotated according to the two dominant orientations are also shown. (Images are generated using the patch extraction system implemented by Bruce Draper.)

where the attention window focuses on a small region and shifts continuously when necessary [123]. Downing and Pinker suggested that attention decreases gradually as the distance from the center of focus increases [40]. Although describing the attention mechanism in different ways, studies show that humans cannot pay attention to more than one set of contiguous locations at the same moment [86]. The goal of our implementation is to provide a way to generate input images for the system.

4.2 Modeling the Primary Visual Cortex

The primary visual cortex (V1) is the first cortical area that receives visual information from the LGN. It consists of about 200 million neurons in total. Since the 1950's, researchers have used single-cell recording techniques to decipher the properties of neurons in V1. Hubel and Wiesel discovered that about 80 percent of the cells in V1 are selective to the orientation of the visual stimulus [64]. They divided the orientation selective cells into two categories, referred as simple and complex cells, based on the receptive field characteristics. Simple cells differ from complex cells in the fact that they show distinct excitatory and inhibitory regions in their receptive fields.

4.2.1 A Simple Cell Model

Since the early visual system is thought to provide the basis for later visual processing, much research effort has been made to seek the receptive field profiles of neurons in V1. Simple cells, the minority of orientation selective cells in V1, respond according to individual spots of light, and the receptive fields of most simple cells have an elongated structure. They are often called edge or line/bar detector cells because they maximally respond to a line/bar or an edge with a particular orientation and retinal location [64, 114]. A simple view of the elongated structure of receptive field and the response profiles of simple cells are shown in Figure 4.2.

Later, more sophisticated simple cell properties were found. At any given visual eccentricity, there are simple cells with different spatial scales of receptive field, i.e. cells optimally tuned to different spatial frequencies [121]. Furthermore, simple cells also respond to stimuli off the preferred orientation and spatial frequencies. In this case, however, the cell response will be lower than the maximal response, which produce additional smaller excitatory and inhibitory waves to the side of the primary peak response (Figure 4.3). How strongly a cell responds to an off-peak stimulus

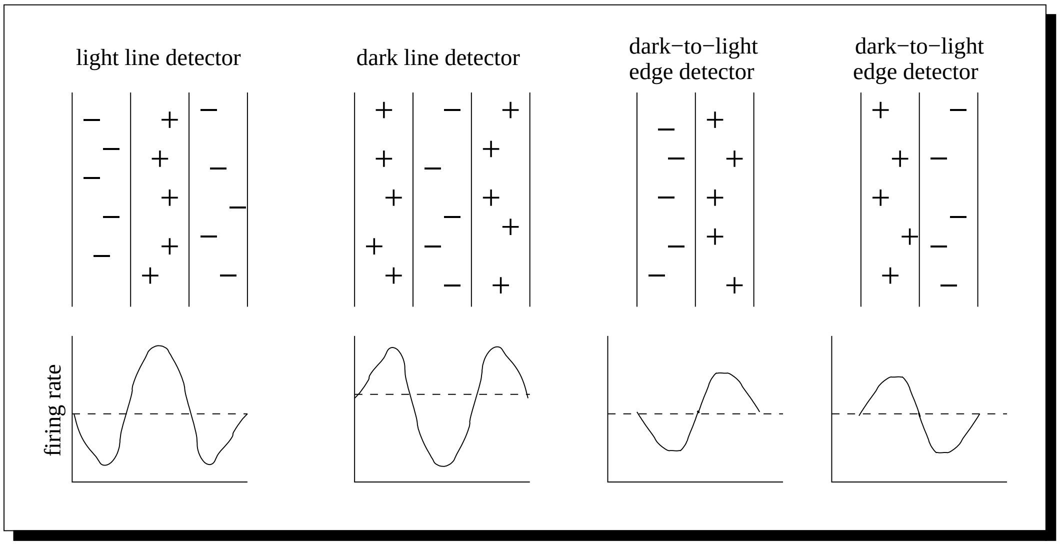


Figure 4.2: Receptive field (top row) and the response profiles (bottom row) of simple cells selective to vertically oriented lines and edges (reprinted from [114]).

depends on the difference between the actual and preferred orientation and spatial frequency, and also on the orientation bandwidth and spatial frequency bandwidth of the cell. The orientation bandwidth of a cell is defined to be the range of stimulus orientations over which the cell response is at least half of its maximal response. Similarly, the spatial frequency bandwidth is defined as the range of spatial frequencies in which the cell response is at least half of its maximal response [116].

The receptive field profiles of simple cells are also considered to be linear in the sense that the response to complicated stimuli can be predicted by the sum of the responses to a set of small stimuli that compose the stimuli [114]. This provides a good reason to model a simple cell as a linear filter with adaptive weights. In fact, it has been shown and generally accepted that the receptive field profiles of simple cells can be approximated by 2D Gabor filter functions, defined as the product of a Gaussian with a sinusoidal wave [116, 121]. (See equation (4.1) below.) Like simple cells, Gabor functions are linear and local, and can be tuned to particular orientations and spatial frequencies. Therefore, the operation of the simple cells in V1 can be modeled

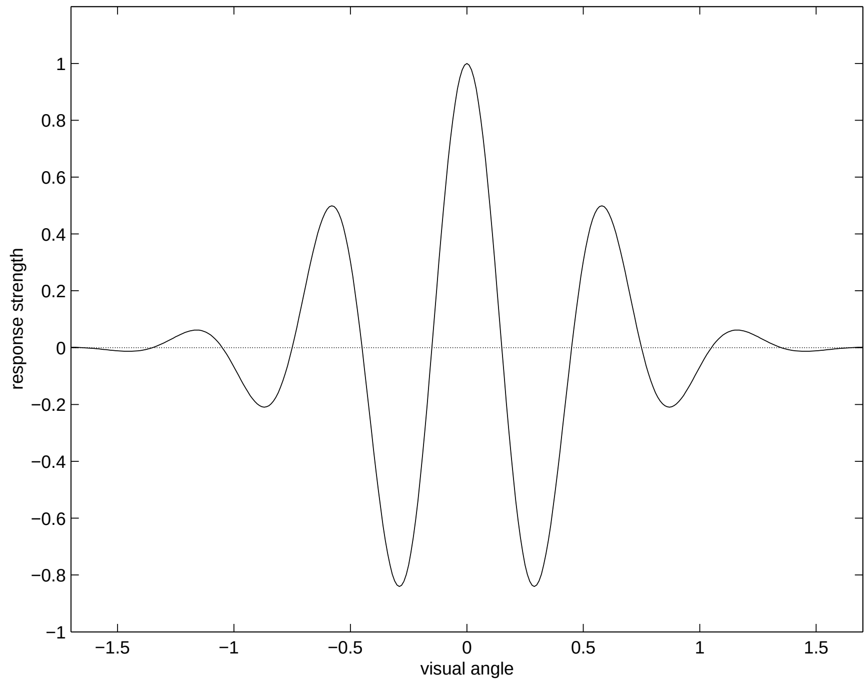


Figure 4.3: 1D view of response profile of a simple cell to a narrow bar in the preferred orientation [114].

as filtering the visual scene through a bank of orientation selective Gabor functions operating at a variety of scales.

A bank of 2D Gabor functions can be defined as follows [88]:

$$g_{\lambda,\theta,\phi} = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\phi\right) \tag{4.1}$$

$$x' = (x,y)\cdot(\cos\theta,\sin\theta) \tag{4.2}$$

$$y' = (x,y)\cdot(-\sin\theta,\cos\theta) \tag{4.3}$$

σ determines the receptive field size and γ, the spatial aspect ratio, specifies the eccentricity of the Gaussian factor. For $\gamma = 1$, the filter is symmetric, and as γ gets smaller, the filter becomes more elongated. λ specifies the wavelength of the cosine factor, and $\frac{\sigma}{\lambda}$ determines the spatial frequency bandwidth b of the filter. b is measured in octaves and defined as follows [88]:

$$b = \log_2 \frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{\ln 2}{2}}}, \quad \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\left(\frac{2^b+1}{2^b-1}\right)$$

Finally, θ specifies the orientation of the filter, and ϕ is the phase offset of the cosine factor. Phase offsets of $0°$ and $180°$ correspond to even symmetric center-on and center-off functions, respectively, while $90°$ and $270°$ correspond to odd symmetric functions. Figure 4.4 shows how the filter changes depending on the parameter values. In this work, the parameters are tuned as suggested in [116] from studies on biological visual systems, with $\gamma = 0.5$ and $\frac{\sigma}{\lambda} = 0.56$.

The patch images extracted by the attention window, as described in Section 4.1, are convolved with a bank of Gabor functions. The resulting responses form a Gabor feature image pyramid for each input image. Figure 4.5 shows sample input patch images and the images of the Gabor filter responses.

4.2.2 A Complex Cell Model

Complex cells make up the majority of neurons in V1 (75%) and are different from simple cells in several ways. Complex cells are highly nonlinear and usually have larger receptive fields than simple cells have. They are more sensitive to motion and invariant to the spatial position of the stimulus in their receptive field, as long as the orientation and direction of motion of the stimulus match the cells' preferred selectivity [114].

The empirical results of complex cell responses are as if the cell sums outputs of several simple cells with the same orientation selectivity, but the phase of their receptive field profiles are off with respect to each other [121]. Pollen and his colleagues describe pairs of simple cells with quadrature or 180 degree inter-pair phase shift relationship while preferred orientation, position, and spatial frequency are maintained [121]. (See the third column of Figure 4.4.) They suggest that these four simple cell

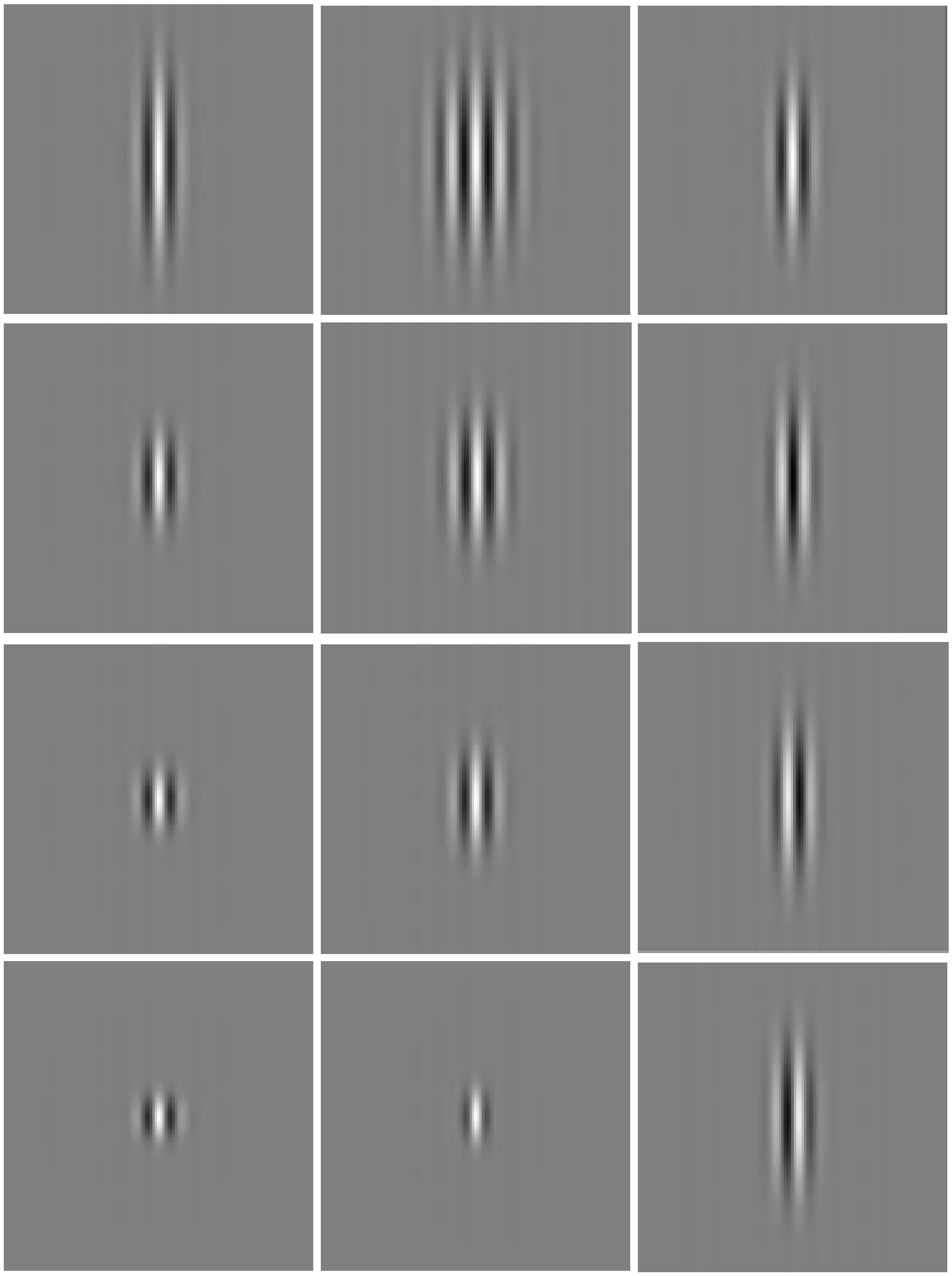


Figure 4.4: 2D Gabor functions. For all, $\lambda = 20$ and $\theta = 0°$. Left column, from top to bottom, $\gamma = 0.25, 0.5, 0.75$, and 1.0 ($\phi = 0$ and $b = 1$). Middle column, from to to bottom, $b = 0.5, 0.7, 0.9$, and 1.8 ($\phi = 0°$ and $\gamma = 0.5$). Right column, from top to bottom, $\phi = 0°, 180°, 90°$, and $270°$ ($\gamma = 0.3$ and $b = 1$). (Images were generated using the applet at [163].)

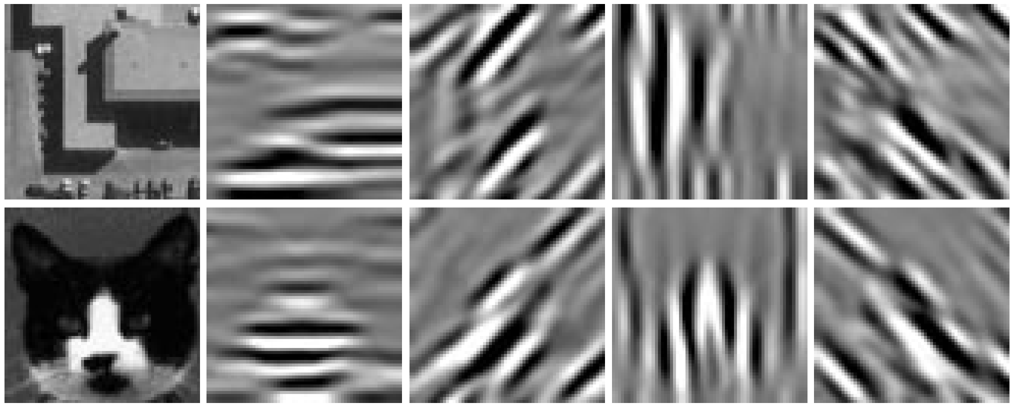


Figure 4.5: Left to right: Input images, filter responses for even symmetric Gabor function with $\theta = 0°$, for odd symmetric Gabor function with $\theta = 45°$, for even symmetric Gabor function with $\theta = 90°$, and for odd symmetric Gabor function with $\theta = 135°$. (Images were generated using the system implemented by Jeff Boody.)

outputs can be used to represent the combined cell responses for the input image. Their complex cell model explains the empirical results quite well and is widely accepted as a general model [88, 116]. In that model, the outputs from simple cells sharing the same orientation selectivity but having 90 degree phase difference are rectified by taking squares (energies), and fed into the complex cells, which then sum them. Computationally, this operation produces the Gabor energy as follows [88]:

$$e_{\lambda,\theta}(x, y) = \sqrt{r^2_{\lambda,\theta,\phi}(x, y) + r^2_{\lambda,\theta,\phi-\frac{1}{2}\pi}(x, y)} \tag{4.4}$$

where $r_{\lambda,\theta,\phi}$ is the image of Gabor responses with the function defined in equation (4.1). Figure 4.6 shows the six Gabor energy images for the cat image shown in Figure 4.5.

4.3 Feature Generation

The visual input changes its representation continuously as it passes through visual areas in the brain. As described in Section 1.1, different cortical areas seem to be largely specialized by their functionality. However, there are few hard biological

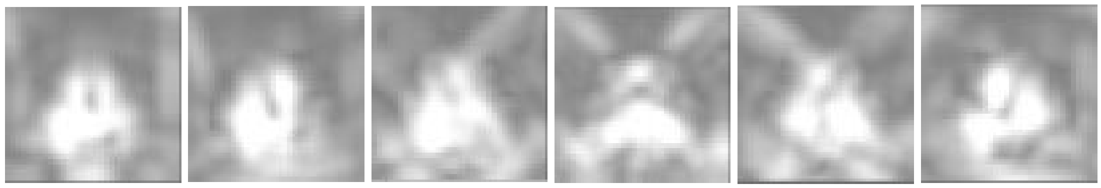


Figure 4.6: Six Gabor energy images computed at a given scale for the Cat image shown in Figure 4.5. From left to right, θ is increased by 15 degrees. The first figure is produced by combining Gabor responses with $\theta = 0$ and 90.

constraints on the actual representation of the visual information, or features, related to the visual processing in each area. For example, there are many different shape features that might be assumed for processing in V2 and many different ways to represent color information for color perception in V4, but no biological evidence shows what and how features are generated when they are needed for processing. Currently, neuronal profiles of representing information in visual areas are not well understood beyond V1. Lacking a definitive model of feature sets, our approach for feature extraction begins with the well-known simple and complex cell models in V1.

The features used in this study are simple features computed using the simple and complex cell responses previously described in terms of Gabor fuctions. These features are basically edge and line-based features; (1) local edge magnitude images are computed from the output of the complex cell model, and (2) linear patterns of local edges are represented in Hough space. There are two reasons for using these simple edge and line-based features. First, biological vision systems are capable of recognizing objects in line drawings. Second, previous results of timing studies for cell activation may restrict the level of complexity of features to be used in expert object recognition.

Using fMRI, Kourtzi & Kanwisher [87] found regions in the lateral occipital complex (LOC) that are activated by both line drawings and grayscale photographs of the same object, as long as the object structures are intact. Furthermore, there is no significant effect from the different stimulus formats on the strength of the activation in those regions. The edge and line features extracted from the grayscale image and line drawings are very similar because the line drawings are usually based on the internal contours and the outlines of the objecs [26]. Therefore, Kourtzi & Kanwisher's work may indicate that such features are used in higher level visual processing. Although LOC seems to respond generally to different types of objects, there is overlap between the LOC and the nearby object-selective regions such as FFA [73]. Thus, we may consider that the regions involved in expert object recognition may also be invariant to the stimulus format, and make use of edges and line features computed in the earlier processing stage.

Expert object recognition is also very fast. While fMRI and PET studies do not give timing information, expert object recognition in humans can also be detected in ERP studies through an early negative component N170, which is found when the subjects perform recognition tasks in their domain of expertise (see Section 3.2). The peak latency of the N170 occurs, on average, 164ms post stimulus onset [140]. This implies that the unique stages of expert object recognition must begin within 164ms of the presentation of the stimulus. The timing studies for the initial responses of the neurons in V1 show that the filtering to the local features appears in about 40 to 80ms post stimulus onset [89]. More complicated features such as texture and boundary completion properties appear as much as 200ms post stimulus onset [89] and, therefore, are probably too late to influence the expert object recognition process.

The features computed in the preprocessing subsystem described by Kosslyn cover a broad range of low-level features computed by the early vision system and more abstract features computed later in the ventral stream. There is, however, no definite resolution showing exactly what features are involved in this stage of processing and how they are computed. Although these features may evolve continuously as they pass through the ventral pathway, biological constraints on the changes of features in relation with the changes of visual processing stages are not well explored at this moment. Therefore, based on the previously described arguments on stimulus format invariance and timing studies, we consider Hough space representation and a combination of complex cell responses as the features computed in the preprocessing stage related to the expert object recognition, and use them in the later processing stages. It should be noted, however, that there are a variety of other features, such as those related to color and shape, that could be added to this processing stage.

4.3.1 Average Complex Cell Edge Magnitude

The edge magnitude features used here are the average complex cell edge magnitudes. They are generated by averaging across orientation the Gabor energies computed by equation (4.4). They are computed at different resolutions to create a pyramid of edge magnitude reponses. Figure 4.7 shows the Gabor energy images and the average edge magnitude image for an input image at a given scale.

4.3.2 Hough Space Representation

The Hough transform is a popular method in image analysis that allows recognition of global patterns in an image space by recognizing local patterns in a transformed parameter space. Generally, it can find complex parameterized features such as straight lines, polynomials, and circles in a suitable parameter space. Here, we use the simplest version of the algorithm to map edges in the input images onto a Hough feature space.

The basic idea is as follows. A straight line in an image space can be represented by the equation $r = x\cos(w) + y\sin(w)$, where r is the perpendicular distance from the origin to the line, and w is the angle of that perpendicular line to the x axis (left graph of Figure 4.8). The Hough space is formed as (r, w) coordinates. As shown in Figure 4.8, the set of lines throuth a point in image space create a curve in Hough

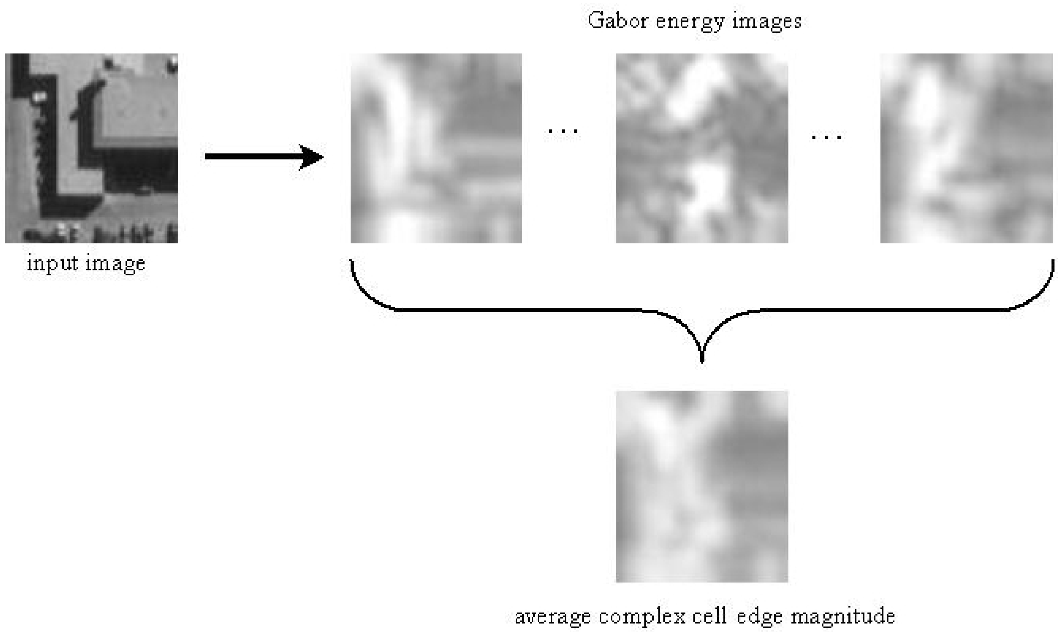


Figure 4.7: Computing average complex cell edge magnitude. The three Gabor energy images are computed using $(0°, 90°)$, $(45°, 135°)$, and $(75°, 165°)$ phase pairs. The average edge magnitude is computed with six Gabor energies computed every 15 degrees. (Images generated using the feature extraction system implemented by Jeff Boody.)

space, and the curves generated by a set of collinear points in image space intersect at the point (r, w) in Hough space, which defines the line in the image space.

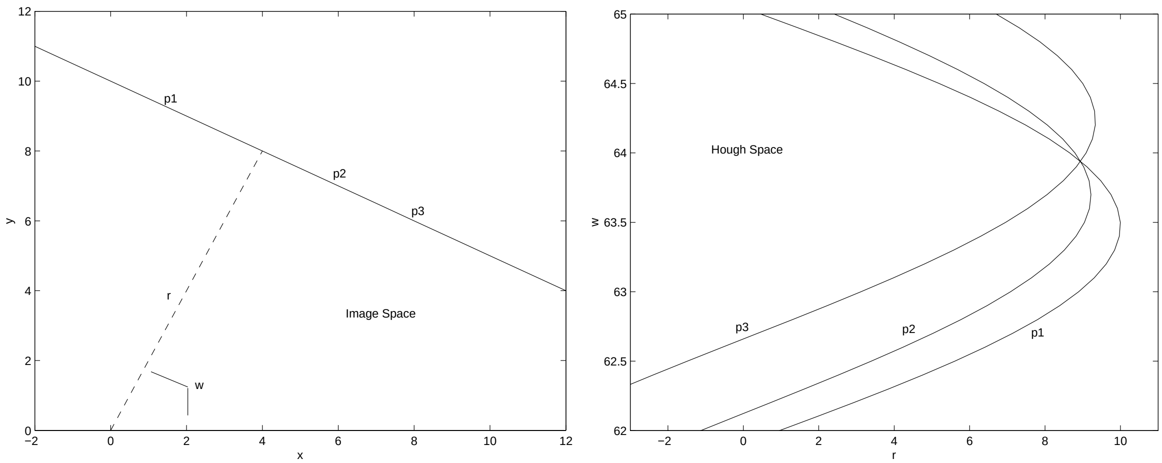


Figure 4.8: The Hough transform. Left figure shows a straight line $y = -0.5x + 10$ in the (x, y) coordinates space, while right figure shows the representation of the three collinear points p1=(1.6, 9.2), p2=(6, 7), and p3=(8, 6) in the Hough space parameterized by r and w. The intersection is approximately $(8.9, 63.9)$.

Hough space can be implemented as a 2D histogram. Each edge pixel in the image provides a radius and orientation of the edge, and is mapped to a point in Hough space; the corresponding bin gets a vote. After all pixels have been processed, the bin with the maximum vote count corresponds to the largest number of collinear edges in the image. Typically, a threshold is applied to determine the salient line segments in the image. Here, the Hough transform is applied to a set of edge images with different orientation but no threshold is applied. The 2D histograms of the Hough space are combined by summing the votes in the corresponding bins and form the final feature vector for the input image. It should be noted that the Hough transform captures collinear and parallel lines in close proximity, therefore implicitly representing some of the non-accidental features described in Section 2.1. Figure 4.9 shows sample Hough feature images.

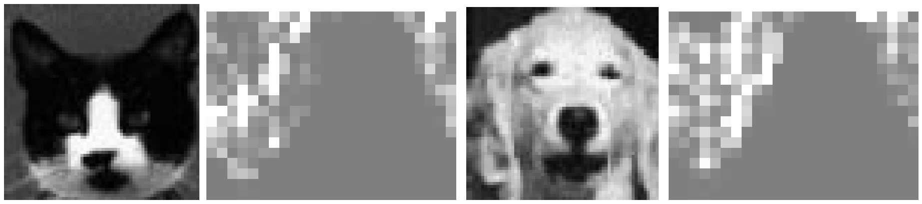


Figure 4.9: The grayscale cat and dog images and their corresponding Hough feature images. In the Hough feature images, vertical axis corresponds to the radius r, and horizontal axis corresponds to the angle w. Origin is the top-left corner.

4.4 Summary

In this chapter, we described the implementation of early stages of the proposed system. It consists of two major components: initial processing for extracting input patches and preprocessing for feature generation. When the system is fed a large image, small patches are extracted around the interest points found by a corner detector. It fills the role of an attention window in biological vision systems, although the actual mechanism does not follow any biological model of attention. Our implementation is just a way of generating input images for the proposed system. Then, the extracted patches are filtered through a bank of orientation selective Gabor funtions operating at different scales. It produces a Gabor feature image pyramid for each patch, which models the output of simple cells in area V1. These features are combined by complex cells to produce a pyramid of phase-insensitive edge/bar energy values.

The representation of visual input evolves continuously through visual areas in the brain. The types of features computed in the preprocessing system described by Kosslyn includes both low-level features computed by the early vision system and more abstract features computed later in visual areas. Less is known about biological constraints on the representation of the visual information related to the processing in visual areas beyond V1. Lacking a definitive model for features, our implementation extracts simple edge and line-based features based on V1 cell responses. We combine complex cell responses into edge magnitude image, and also use Hough space representation over edges in the input image. Our use of this simple feature set is based on an fMRI study showing that edge-based features are indeed used in expert object recognition system, and the timing studies for cell activation indicating that more complicated features are probably too late to influence the expert object recognition process.

In the next chapter, we will present the implementation of the pattern matching system, which consists of two separate processes: category and exemplar recognition. We will describe a set of unsupervised clustering and subspace projection algorithms as computational methods for the two stages processing.

Chapter 5

Pattern Matching

The pattern matching stage consists of two separate processes. The classification process corresponds to the category pattern activation subsystem and the exemplar recognition process corresponds to the exemplar pattern activation subsystem. The goals of the two processes differ in terms of the level of recognition; classification is *basic* or *superordinate* level recognition, while exemplar recognition works at the *subordinate* level. Although Poggio & Hurlbert [120] proposed an architecture for recognition with both functionalities, as they also mentioned, most object recognition systems tackle the tasks separately; they are designed to detect a certain type of object (e.g. face) in an image cluttered with other objects (classification), or to identify individual instances presented in the image (exemplar recognition). In this work, we implement both level of tasks in the proposed system and make hierarchical interconnections between them. It should be noted, however, that there may also be direct routes from early stages where features are extracted to the exemplar pattern recognition subsystem.

As discussed in Chapter 3, the category and exemplar pattern activation subsystems are associated with different receptive field sizes, and therefore use different types of information to achieve their goal. This motivates us to propose that different mechanisms are used to implement each subsystem. Classification is modeled as unsupervised clustering during training followed by maximum likelihood classification during testing, and exemplar recognition is implemented as subspace extraction during training and subspace projection followed by nearest neighbor matching during testing. In this work, a unique subspace is computed for every cluster defined by the category subsystem, therefore, a set of local linear models is built rather than a single, global model as in many PCA-based systems. The local linear approaches may model the data containing multiple classes of objects better than global approaches. Localized models have previously been used for face recognition [47] and image compression and reconstruction [75], but not in the context of multiple object classes. The argument is that global PCA subspaces are optimal when the images are drawn from a single underlying normal distribution, however, they may not align to any meaningful axis of variance if the underlying distribution is a mixture of normal distributions. A local subspace for each population is required for this case. The argument is even stronger in the context of expert object recognition, since people are experts in recognizing many types of objects, and images of different objects are unlikely to come from a single normal distribution.

5.1 Classification

The task of classification is to assign images to categories. In this work, categories are not linguistic or logical labels that share semantic properties but groups of images that are visually similar. It is the task of super-visual processes accessing associative memory to assign linguistic or logical labels to visual categories, not the task of the ventral visual pathway. Therefore, classification is modeled by unsupervised clustering algorithms operating on previously computed feature values.

There exist many algorithms for unsupervised clustering. In this study, we investigate three algorithms: K-Means, EM with a mixture of Gaussians, and clustering with probabilistically weighted PCA.

5.1.1 K-Means Clustering

The K-Means algorithm [42] partitions the input samples into K disjoint clusters such that all data points in a cluster are associated with the center (or prototype) of the cluster. K-Means starts by initializing the K centers to either random values or the positions of randomly chosen training samples. Then, it keeps track of the centers of the clusters as it proceeds through iterations performing the following two steps:

1. Classify the data: For each training sample, find the closest center and assign the sample to the center. If two or more centers are equally close to a sample, break the ties by random selection.

2. Update the K centers: The center of each cluster is recomputed by taking the mean of all samples assigned to it.

The algorithm terminates when no sample changes clusters, or there is no significant changes in the center locations, or after running a fixed number of iterations.

K-Means is a very simple and robust unsupervised clustering method. It is known that K-Means converges to a local optimum which minimizes the sum of squared distances between data samples in a cluster and the cluster's center [42]. It is essentially an approximation to a maximum-likelihood estimates for cluster centers, under the assumption that every data cluster represents a symmetric Gaussian distribution located at the cluster center.

5.1.2 Mixture of Gaussians

A more general mixture model framework can effectively represent arbitrarily shaped clusters by selecting appropriate component density functions. Mixture models assume that the observed data set is a mixture of samples from multiple populations. The probabilistic nature of mixture models do not require the specification of a distance metric, and allows soft decision criteria as oppose to the hard decision made by K-Means clustering.

In a mixture model, component density functions are selected, and then the unknown parameters are estimated. One of the classic and popular techniques used for mixture model parameter estimation is the *Expectation-Maximization* (EM) algorithm [25, 35]. EM finds the maximum-likelihood estimate of the parameters of a mixture model. Let $\mathbf{g}_i$, $i = 1, \ldots, K$, be K component density functions and $\mathbf{x} \in R^p$ be a data sample. Then, the mixture model probability density function evaluated at $\mathbf{x}$ is:

$$G(\mathbf{x}|\mathbf{\Psi}) = \textstyle\sum_{i=1}^{K} w_i \mathbf{g}_i(\mathbf{x}|\theta_i)$$

where $\mathbf{\Psi} = (w_i, \theta_i)$ is a set of parameters and w_i are weights for corresponding clusters. Thus, $\sum_i w_i = 1$ and $w_i \geq 0$.

The log-likelihood expression of the data set given the density functions is:

$$\begin{aligned} L(\mathbf{\Psi}) &= \log \prod_{i=1}^{M} G(\mathbf{x}_i|\mathbf{\Psi}) \\ &= \sum_{i=1}^{M} \log G(\mathbf{x}_i|\mathbf{\Psi}) \\ &= \sum_{i=1}^{M} \log \left(\sum_{h=1}^{K} w_h \mathbf{g}_h(\mathbf{x}_i|\theta_h) \right) \end{aligned}$$

where M is the number of data samples. The log-likelihood quantifies the quality of the set of parameters $\mathbf{\Psi}$, which explains how well the mixture model fits the data set. The EM algorithm finds $\mathbf{\Psi}$ such that $L(\mathbf{\Psi})$ is maximized.

In this work, we model each cluster by a multivariate Gaussian probability distribution. The multivariate Gaussian for cluster $i = 1, \ldots, K$, is parameterized by the mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$ (i.e. $\mathbf{\Psi} = (w_i, \mu_i, \mathbf{\Sigma}_i), i = 1, \ldots, K$). Then,

$$\mathbf{g}_i(\mathbf{x}|\theta_i) = \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma}_i)}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \mu_i) \right\}$$

EM begins by initializing parameters, μ_i, $\mathbf{\Sigma}_i$, and w_i for $i = 1 \ldots K$, and iterates the following two steps:

- E-step: Estimate the membership probability of $\mathbf{x}$ for each cluster i

$$P_i(\mathbf{x}) = \frac{w_i \mathbf{g}_i(\mathbf{x})}{\sum_{i=1}^{K} w_i \mathbf{g}_i(\mathbf{x})}$$

- M-step: Update mixture model parameters. The log-likelihood computed using the newly estimated parameters increases from the previous iteration.

$$\begin{aligned}
w_i^{new} &= \frac{\sum_{\mathbf{x}} P_i(\mathbf{x})}{M} \\
\mu_i^{new} &= \frac{\sum_{\mathbf{x}} P_i(\mathbf{x})\mathbf{x}}{\sum_{\mathbf{x}} P_i(\mathbf{x})} \\
\mathbf{\Sigma}_i^{new} &= \frac{\sum_{\mathbf{x}} P_i(\mathbf{x})\{(\mathbf{x} - \mu_i^{new})(\mathbf{x} - \mu_i^{new})^T\}}{\sum_{\mathbf{x}} P_i(\mathbf{x}) - 1}
\end{aligned}$$

The EM algorithm terminates if $|L(\mathbf{\Psi}) - L(\mathbf{\Psi}^{new})| \leq \epsilon$ or the number of iterations reaches to the given maximum value.

Clustering by fitting Gaussian mixtures using the EM algorithm allows for asymmetric Gaussian distributions to model each cluster. However, when the dimension of data samples gets large, it suffers from numerical instability (including underflow) and a possibly singular covariance matrix to be inverted when computing the probability. Therefore, it is difficult in practice to apply EM with Gaussian mixture model to high dimensional data.

5.1.3 Clustering with Probabilistically Weighted PCA

To fit asymmetric Gaussians while avoiding these problems, we approximate the multi-dimensional Gaussian by decomposing it into two parts using PCA. The decomposition of the probability function can be shown as in [106]. Let us assume that the density function of input data $\mathbf{x}$ is a multivariate p-dimensional Gaussian:

$$p(\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\bar{\mathbf{x}})\right\}}{2\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}}$$

where $\bar{\mathbf{x}}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the training set $\{\mathbf{x}_i|\mathbf{x}_i \in R^p, i = 1, \ldots, n\}$. As described in Section 5.3.1, PCA factors $\boldsymbol{\Sigma}$ using an orthogonal matrix $\boldsymbol{\Upsilon}$ of eigenvectors and a diagonal matrix $\boldsymbol{\Lambda}$ of eigenvalues λ_i's:

$$\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}\boldsymbol{\Lambda}\boldsymbol{\Upsilon}^T$$

Then, the Mahalanobis distance term can be represented as follows:

$$\begin{aligned}
(\mathbf{x}-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) &= \tilde{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{x}} \\
&= \tilde{\mathbf{x}}^T(\boldsymbol{\Upsilon}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Upsilon}^T)\tilde{\mathbf{x}} \\
&= \mathbf{z}^T\boldsymbol{\Lambda}^{-1}\mathbf{z} \\
&= \sum_{i=1}^{p} z_i^2/\lambda_i \\
&= \sum_{i=1}^{q} z_i^2/\lambda_i + \sum_{i=q+1}^{p} z_i^2/\lambda_i
\end{aligned}$$

where $\mathbf{z}$ is the principal components vector obtained by $\mathbf{z} = \boldsymbol{\Upsilon}^T\tilde{\mathbf{x}}$. The first part corresponds to the error in the q-dimensional principal subspace and the second part is the error in the orthogonal complementary subspace (reconstruction error).

Therefore,

$$\begin{aligned}
p(\mathbf{x}) &= \frac{\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i} + \sum_{i=q+1}^{p}\frac{z_i^2}{\lambda_i}\right)\right\}}{2\pi^{p/2}|\Sigma|^{1/2}} \\
&\approx \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i}\right\}}{(2\pi)^{q/2}\prod_{i=1}^{q}\lambda_i^{1/2}} \times \frac{\exp\left\{-\frac{1}{2\sigma}\sum_{i=q+1}^{p} z_i^2\right\}}{(2\pi\sigma)^{(p-q)/2}}
\end{aligned}$$

where σ is the average of the trailing eigenvalues $\lambda_{q+1}, \ldots, \lambda_p$. $p(\mathbf{x})$ can be computed using only the q-dimensional principal components since the error term $\sum_{i=q+1}^{p} z_i^2$ is equal to $\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2$. It makes

$$p(\mathbf{x}) \approx \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i}\right\}}{(2\pi)^{q/2}\prod_{i=1}^{q}\lambda_i^{1/2}} \times \frac{\exp\left\{-\frac{1}{2\sigma}\left(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2\right)\right\}}{(2\pi\sigma)^{(p-q)/2}} \tag{5.1}$$

This density estimation using PCA subspace decomposition has been applied to probabilistic target detection and object recognition [106].

The clustering algorithm assigns a sample to each cluster with a probability computed as in equation (5.1). Then, a new PCA subspace is formed independently for each cluster taking the data samples weighted by the probability as input. The *probabilistically weighted* PCA (PWPCA) clustering algorithm proceeds as follows:

1. For each cluster C_j ($j = 1, \ldots, k$), initialize the probability of $\mathbf{x}_i$ drawn from cluster C_j, $p(\mathbf{x}_i|C_j)$, using Euclidean distance to the randomly selected cluster center μ_j. Update μ_j as weighted mean: $\frac{1}{\sum_i w_{ij}} \sum_i w_{ij}\mathbf{x}_i$, where $w_{ij} = \frac{p(\mathbf{x}_i|C_j)}{\sum_j p(\mathbf{x}_i|C_j)}$.

2. Perform PCA for C_j on the weighted input, $w_{1j}(\mathbf{x}_1 - \mu_j), \ldots, w_{nj}(\mathbf{x}_n - \mu_j)$.

3. Recompute $p(\mathbf{x}_i|C_j)$ using newly formed PCA subspace and update μ_j's and w_{ij}'s.

4. Iterate step 2 and 3 until it stabilizes.

5. For each sample $\mathbf{x}_i$, assign it to the maximum-likelihood cluster:

$$\mathbf{x}_i \in C_j, \text{ if } p(\mathbf{x}_i|C_j) > p(\mathbf{x}_i|C_l), \ \forall l \neq j$$

PWPCA clustering is basically a combination of nearest neighbor partitioning by the sample probability and probability computation using PCA. A similar approach using nearest neighbor clustering and PCA is also proposed by Kambhatla and Leen [75]. Unlike PWPCA, their algorithm, called VQPCA, utilizes only the reconstruction error as the clustering criterion since VQPCA is applied to the dimensionality reduction problem. Another difference is that, while PWPCA softly assigns each sample to all clusters according to the probability, VQPCA uses hard assignment during clustering in each iteration.

The PWPCA clustering algorithm is intimately related to Tipping and Bishop's *Mixture of Probabilistic* PCA model [148]. They introduced a probabilistic model for PCA and developed an EM algorithm for a mixture model of principal component analyzers. It iteratively estimates maximum-likelihood model parameters and partitions the data set according to the mixture component's responsibility for generating each data sample. As shown in [148], the responsibility is proportional to the probability given in equation (5.1), therefore, we believe[1] that it operates similarly to PWPCA clustering. Tipping and Bishop's approach also avoids the dimensionality problem that the traditional EM for Gaussian mixtures has, requiring the inversion of a $q \times q$ matrix instead of a $p \times p$ matrix. The PWPCA clustering algorithm which we developed uses conventional closed-form PCA and computes the probability using eigenvalues and subspace projection vectors. A numerical stability problem can occur for probability computation in both algorithms due to the large negative exponential values. The implementation details that avoid this problem are described in Appendix B.

5.2 Illustration of Clustering Algorithms for Synthetic Data

In this section, we briefly show how the three clustering algorithms work using a 2D synthetic data set, which consists of a mixture of two Gaussian distributions.

5.2.1 Data Set

To illustrate the clustering algorithms, we generated a 2D synthetic data set with 800 data points randomly sampled from two Gaussian distributions. To generate Gaus-

[1]At this moment, we are not able to directly compare the clustering results from both algorithms.

sian random points, we used the *Box-Muller* method, which transforms uniformly distributed random variables to a new set of random variables with a Gaussian distribution [124]. Let x_1 and x_2 be random variables drawn from a uniform distribution in (0, 1). Then, the following transformation produces two new random variables y_1 and y_2, which have a Gaussian distribution with zero mean and a standard deviation of one.

$$\begin{aligned} y_1 &= \sqrt{-2\ln x_1}\cos(2\pi x_2) \\ y_2 &= \sqrt{-2\ln x_1}\sin(2\pi x_2) \end{aligned}$$

If we set

$$\begin{aligned} z_1 &= s_1 y_1 + m_1 \\ z_2 &= s_2 y_2 + m_2 \end{aligned}$$

then, z_1 and z_2 are random variables drawn from $\mathcal{N}(m_1, s_1)$ and $\mathcal{N}(m_2, s_2)$, respectively. Therefore, (z_1, z_2) forms a 2D Gaussian random point. In this way, we generated 800 data points from two different Gaussian distributions with means of (0, 0) and (15, -7), and standard deviations of (12, 4) and (2, 4). Then, we rotated the points 45 degrees and shift the rotated points by (45, 45). Rotation of the data points makes the standard deviations for the x and y dimensions no longer independent. Figure 5.1 shows the resulting data set. The mean locations of the two Gaussian distributions are (45, 45) and (60, 50).

5.2.2 Clustering Results

We tested K-Means, traditional EM, and PWPCA clustering algorithms with two clusters for ten iterations. The clustering progress for each algorithm is shown in Figure 5.2 and Figure 5.3. Note that the scales are different for PWPCA because the points are scaled between 0 and 255 in order to use the OpenCV eigen-decomposition

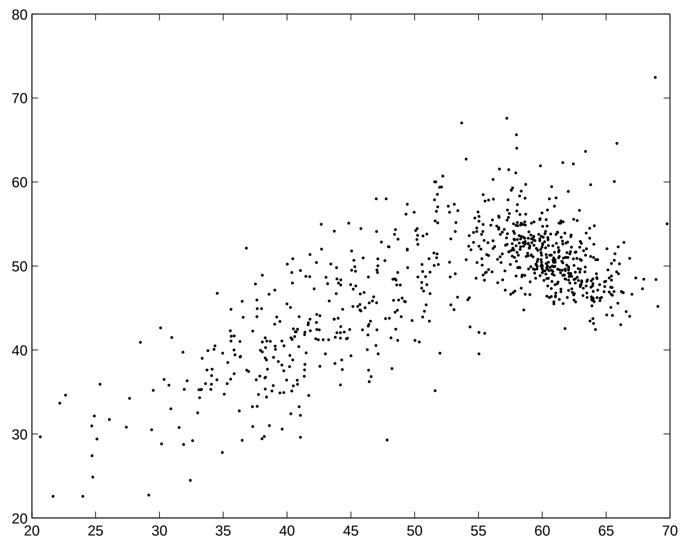


Figure 5.1: The 2D synthetic data set.

function. Although soft assignment is used in each iteration for traditional EM and PWPCA clustering, the resulting clusters in the figures were formed by hard assignment; a point is assigned to the cluster with higher probability. As we can see in the figures, K-Means clusters look more like a mixture of circularly symmetric Gaussian distributions, while the clusters from the other two algorithms fit the true asymmetric Gaussian distributions more closely. The mean locations that K-Means found are (40, 40) and (59, 51) while traditional EM finds clustering centers (45, 45) and (61, 51). The means found by PWPCA clustering are (116, 117) and (193, 148), which correspond to (44, 44) and (60, 51) in the original space. The mean locations found by traditional EM and PWPCA clustering are closer to the true means, (45, 45) and (60, 50), than those found by K-Means.

To show the behavior of PWPCA clustering in more detail, we plot the principal axis for each cluster in Figure 5.4. The data points are computed by subtracting weighted-mean of the training samples and then weighting them according to the probability for the cluster. In the figures at the bottom row of Figure 5.3, the small group of points in the middle-left side that are clustered differently from the sur-

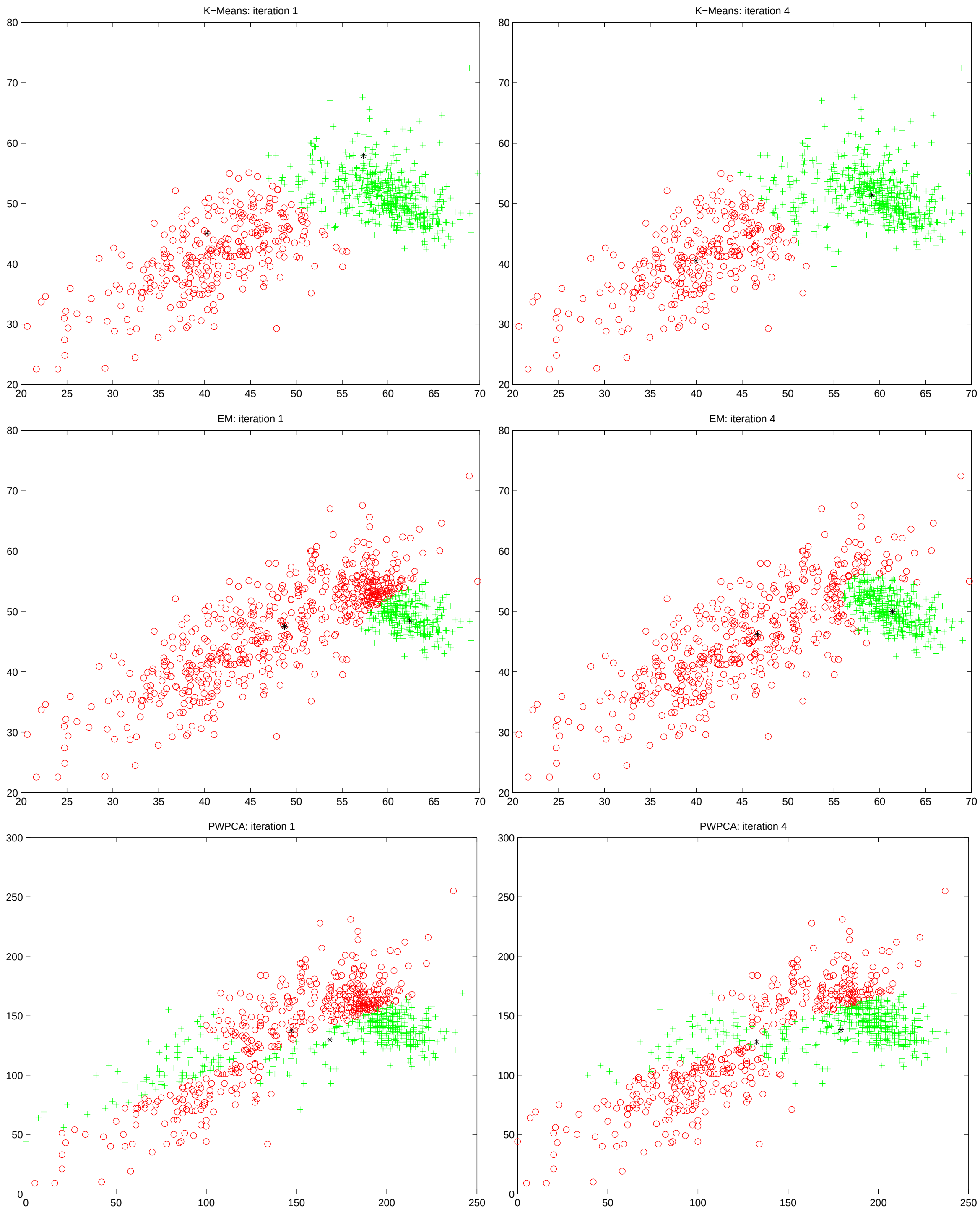


Figure 5.2: Intermediate clustering results at iteration 1 (left) and iteration 4 (right) for K-Means (top), traditional EM (middle), and PWPCA clustering (bottom). Star (*) is the cluster mean.

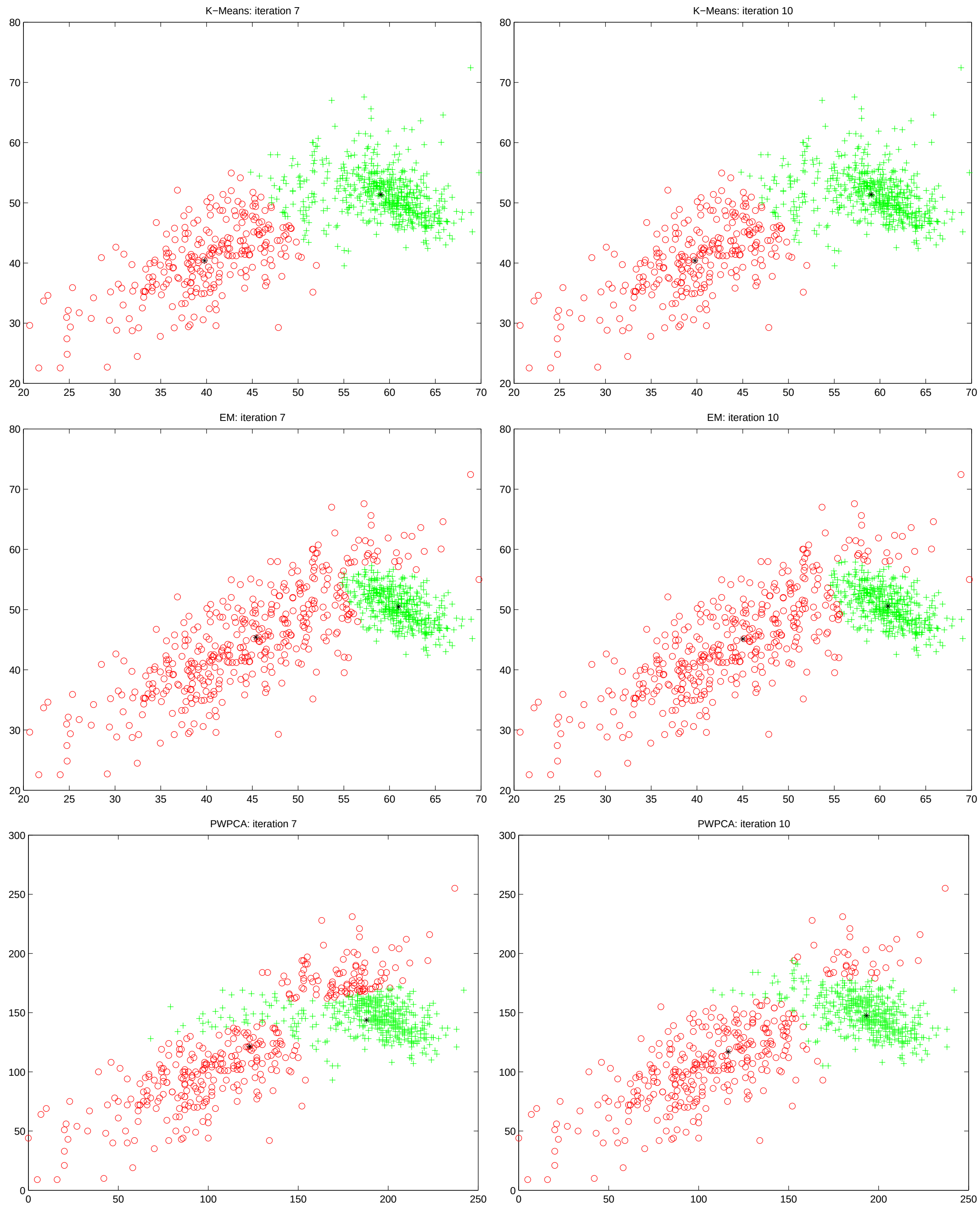


Figure 5.3: Intermediate clustering results at iteration 7 (left) and iteration 10 (right) for K-Means (top), traditional EM (middle), and PWPCA clustering (bottom). Star (*) is the cluster mean.

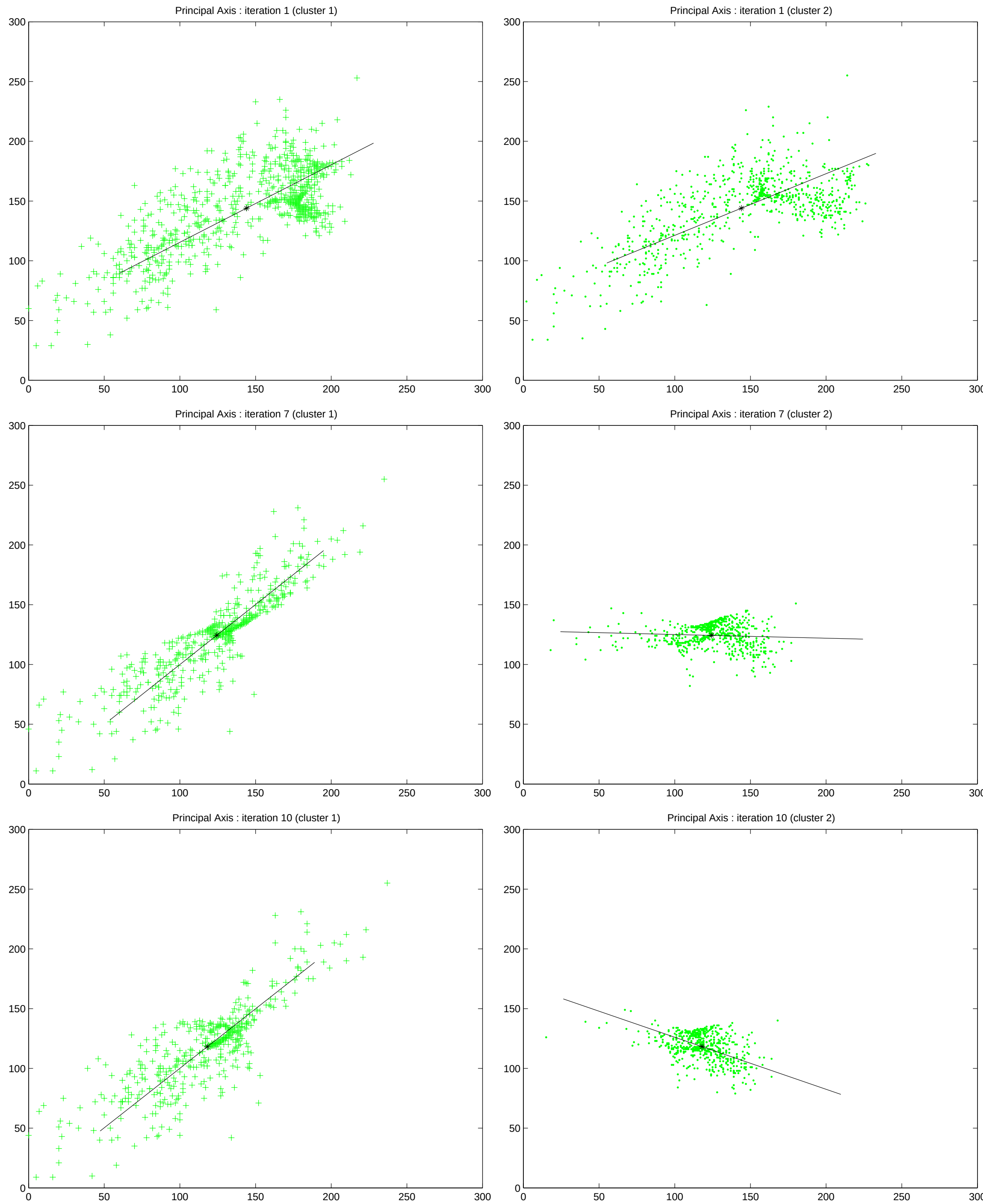


Figure 5.4: Principal axis computed by PWPCA for cluster 1 (left) and cluster 2 (right) at iteration 1 (top), 7 (middle), and 10 (bottom). The data points are weighed mean-subtracted values.

rounding points is the result of taking both principal subspace and outer-subspace distance into account for probability computation. For comparison, Figure 5.5 shows the resulting clusters obtained by using the outer-subspace distance (reconstruction error) only.

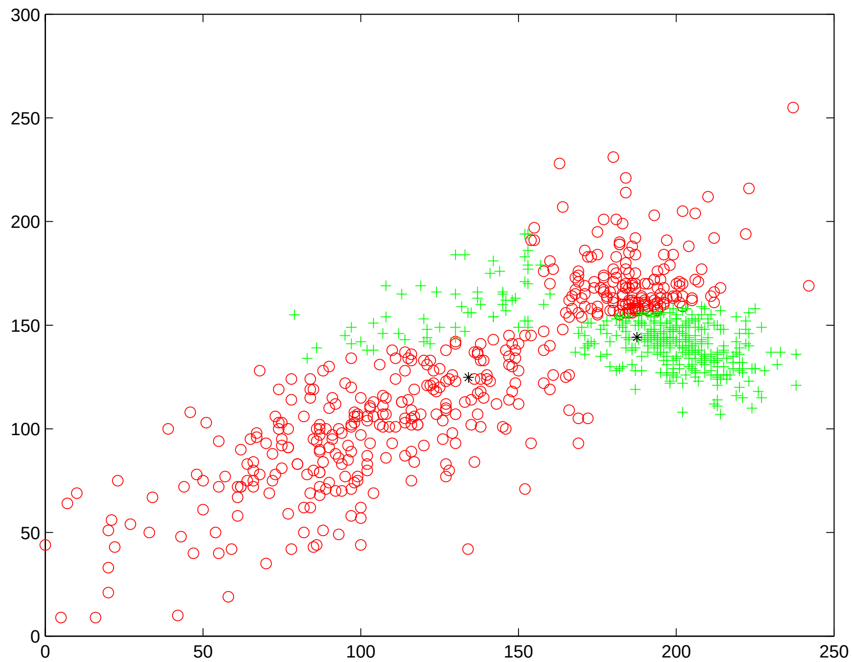


Figure 5.5: PWPCA clustering result obtained when only the reconstruction error was used as a clustering criterion.

5.3 Exemplar Recognition

While the classification system matches images to categories – clusters of images in memory – the exemplar recognition system matches current input images to a specific image in visual memory. As for the actual matching mechanism, Kosslyn denies that a complete template-like image is generated to match to the input, partly because no anatomical structure matching a template has ever been found, and partly because template matching is considered as too rigid for biological vision. Instead, he suggests that objects are represented as "compressed images" in the visual memory and used by the exemplar subsystem. Compressed images are smaller than the source images, and yet sufficient to regenerate an approximation of the source image for imagery feedback.

The notion of "compressed images" can be modeled as subspace projection for image matching in the machine vision literature. Subspace projection techniques define the input images as points in a p-dimensional space where p is the number of pixels in each image. p is generally very large and the meaningful features are often obscured by noise and complicated dependencies between variables in the p-dimensional space. Therefore, it may be important to reduce the dimensionality of the input data by projecting it into a smaller and more manageable space in which the relevant features are more explicit. In many cases, the mapping is sought as a linear transformation of $\mathbf{x} \in R^p$ into $\mathbf{z} \in R^q$ where $q < p$, i.e.

$$\mathbf{z} = \mathbf{W}\mathbf{x}$$

The projections, $\mathbf{z}$, are the "compressed images" referred to by Kosslyn. New images are projected into the subspace and matched by measuring the subspace distance from previously stored images. In this work, we have considered three different subspace projection algorithms for exemplar recognition: principal component analysis (PCA), independent component analysis (ICA), and factor analysis (FA).

5.3.1 PCA

PCA has been widely used for object recognition since it was first applied to face recognition [79, 151]. PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. Let $\mathbf{w}_1$ be the direction of the first principal component. Then $\mathbf{w}_1$ is defined as

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} E[(\mathbf{w}^T\mathbf{x})^2]$$

Thus the first principal component is the projection of the data $\mathbf{w}^T\mathbf{x}$ onto the direction that maximize the variance of the projection. The i-th principal component is found by

$$\mathbf{w}_i = \arg\max_{\|\mathbf{w}\|=1} E[\{\mathbf{w}^T(\mathbf{x} - \sum_{c=1}^{i-1} \mathbf{w}_c\mathbf{w}_c^T\mathbf{x})\}^2]$$

This results in a set of orthonormal (linearly uncorrelated) vectors, $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$.

In practice, the eigenvectors of the sample covariance matrix of the data set are used to compute the principal components. Let $\mathbf{\Sigma}$ be the sample covariance matrix. Then, the eigenvalue decomposition of $\mathbf{\Sigma}$ is

$$\mathbf{\Upsilon}^T\mathbf{\Sigma}\mathbf{\Upsilon} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, and $\mathbf{\Upsilon} = [\mathbf{w}_1|\mathbf{w}_2|\cdots|\mathbf{w}_q]$ is an orthonormal matrix whose columns are the q principal eigenvectors of $\mathbf{\Sigma}$. Geometrically, $\mathbf{\Upsilon}$ rotates the coordinate system onto the eigenvectors, where the eigenvector associated with the largest eigenvalue is the axis of maximum variance, the eigenvector associated with the second largest eigenvalue is the axis with the second largest variance orthogonal to the previous eigenvector, etc.

The variables $\mathbf{z} = (z_1, \ldots, z_q)$ computed by the transformation $\mathbf{z} = \mathbf{\Upsilon}^T\mathbf{x}$ are the principal components of the input vector $\mathbf{x}$. The principal components are uncorrelated, i.e. $cov(z_i, z_j) = 0$ for all $i \neq j$, and the variance of z_i is λ_i. Therefore, the first principal component z_1 is the linear combination of $x_1, \ldots, x_p$ with the highest variance and, similarly, z_i has the highest variance among all linear combinations of $x_1, \ldots, x_p$ which are uncorrelated with $z_1, \ldots, z_{i-1}$. In other words, each principal component captures 'as much as possible' of the variation in $\mathbf{x}$ unexplained by previous principal components. The number of principal components, q, is selected so that the first q principal components explain enough of the total variation. This mapping of $\mathbf{x}$ to a lower dimensional representation $\mathbf{z}$ is optimal in the mean squared error sense. That is, the inverse mapping of $\mathbf{z}$ back into $\mathbf{x}$ has minimum reconstruction error. In fact, if $\boldsymbol{\Upsilon}$ includes all the eigenvectors with non-zero eigenvalues, the inverse mapping is lossless.

Although PCA can be implemented by standard closed-form numerical approaches performing the eigen-decomposition directly from the input matrix, there also exist biologically inspired neural network algorithms [118, 130]. They use a linear neural network with a layer of neurons that receive their input via Hebbian synaptic connections, and an anti-Hebbian learning rule for the lateral connection within the output layer. The weight vectors learned by the neural network tend to converge to the eigenvectors of the data correlation matrix. In this study, however, we use a standard closed-form method for implementing PCA.

5.3.2 ICA

While PCA decorrelates the input data using second-order statistics and thereby generates compressed data with minimum mean-squared reprojection error, ICA minimizes higher-order dependencies in the input (in addition to linear decorrelation). It is intimately related to the blind source separation (BSS) problem, where the goal is to recover the unknown independent signals from observed linear combinations of them. Let $\mathbf{z}$ be the vector of unknown source signals and $\mathbf{x}$ be the vector of observed mixtures. If $\mathbf{A}$ is the unknown mixing matrix that relates the observed and the source signals, then the mixing model is written as

$$\mathbf{x} = \mathbf{A}\mathbf{z}$$

It is assumed that the source signals are independent of each other and the mixing matrix $\mathbf{A}$ is invertible. Based on these assumptions and the observed mixtures, ICA algorithms try to find the mixing matrix $\mathbf{A}$ or the separating matrix $\mathbf{W}$ such that

$$\mathbf{u} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{z}$$

is an estimation of the independent source signals [31] (Figure 5.6).

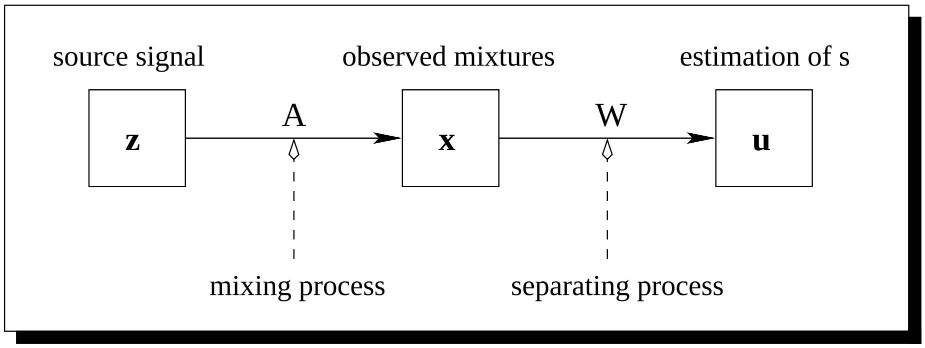


Figure 5.6: Blind source separation model.

ICA can be viewed as a generalization of PCA. As previously discussed, PCA decorrelates the data so that the covariance of the data is zero. A slightly stronger constraint than being uncorrelated is whiteness, which further restricts the data to have unit variance. The whitening transform can be determined as $\mathbf{D}^{-1/2}\mathbf{R}^T$ where $\mathbf{D}$ is the diagonal matrix of the eigenvalues and $\mathbf{R}$ is the orthogonal matrix of eigenvectors of the covariance matrix. Applying whitening to observed mixtures, however, results in the source signal only up to an orthogonal transformation. ICA goes one step further so that it transforms the whitened data into a set of statistically independent signals [68].

Unlike PCA, there is no closed form expression to find $\mathbf{W}$. Instead, many iterative algorithms have been proposed based on different search criteria [69]. However, it has been shown that most of the criteria optimized by different ICA algorithms lead to similar or even identical algorithms [31, 67]. In this work, we will use *InfoMax*, one of the best-known ICA algorithms by Bell and Sejnowski [14]. InfoMax is a gradient based, unsupervised learning algorithm for ICA based on the information maximization principle of data transfer between sigmoidal neurons. The information maximization criterion is essentially equivalent to the maximum likelihood criterion [31].

5.3.2.1 Architecture I: Statistically Independent Basis Images

In [9], Bartlett et al. introduced two different ways to apply ICA to object recognition. In Architecture I, the input images in **X** are considered to be a linear mixture of statistically independent basis images in **Z** combined by an unknown mixing matrix **A**. The InfoMax algorithm learns the weight matrix **W**, which is used to recover a set of independent basis images in the rows of **U** (Figure 5.7). In this architecture, the input images are variables and the pixel values provide observations for the variables. The source separation, therefore, is performed in image space and finds the weight vectors in **W** in the directions of statistical dependencies in the set of input images. The independent basis images, which span the image space, are produced by projecting the input images onto the learned weight vectors. The compressed representation of an image is a vector of components used for linearly combining the independent basis images to generate the image. The middle row of Figure 5.8 shows eight basis images produced in this architecture. They are spatially localized, unlike the PCA basis images (top row) and those produced via Architecture II (bottom row).

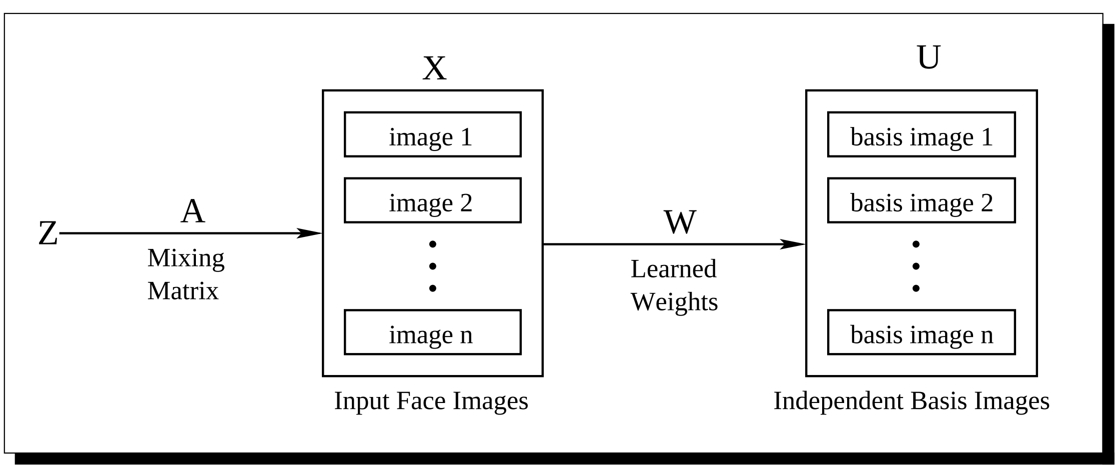


Figure 5.7: Finding statistically independent basis images.

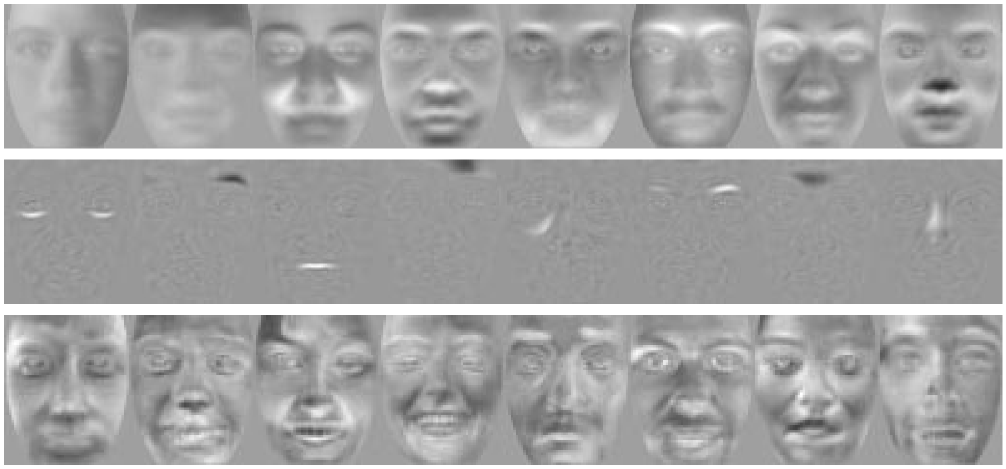


Figure 5.8: Eight basis vectors for PCA and ICA computed on a face image data set. The top row contains the eight eigenvectors with highest eigenvalues for PCA. The second row shows eight localized basis vectors for ICA Architecture I. The third row shows eight, non-localized ICA basis vectors for ICA Architecture II.

In [9, 10], Bartlett and colleagues first apply PCA to project the data into a subspace of dimension q to control the number of independent components produced by ICA. The InfoMax algorithm is then applied to the eigenvectors to minimize the statistical dependence among the resulting basis images. This use of PCA as a pre-processor in a two-step process allows ICA to create subspaces of size q for any q. In [90], it is also argued that pre-applying PCA enhances ICA performance by (1) discarding small trailing eigenvalues before whitening and (2) reducing computational complexity by minimizing pair-wise dependencies. PCA linearly decorrelates the input data. The remaining higher-order dependencies are separated by ICA.

5.3.2.2 Architecture II: Statistically Independent Components

While the basis images obtained in Architecture I are statistically independent, the components that represent input images in the subspace defined by the basis images are not. The goal of ICA in Architecture II is to find statistically independent components for input data. In this architecture, the input is transposed from architecture I, that is, the pixels are variables and the images are observation. The source separation is performed on the pixels, and each row of the learned weight matrix **W** is an image. **A**, the inverse matrix of **W**, contains the basis images in its columns. The statistically independent source components in **Z** that comprise the input images are recovered in the columns of **U** in Figure 5.9.

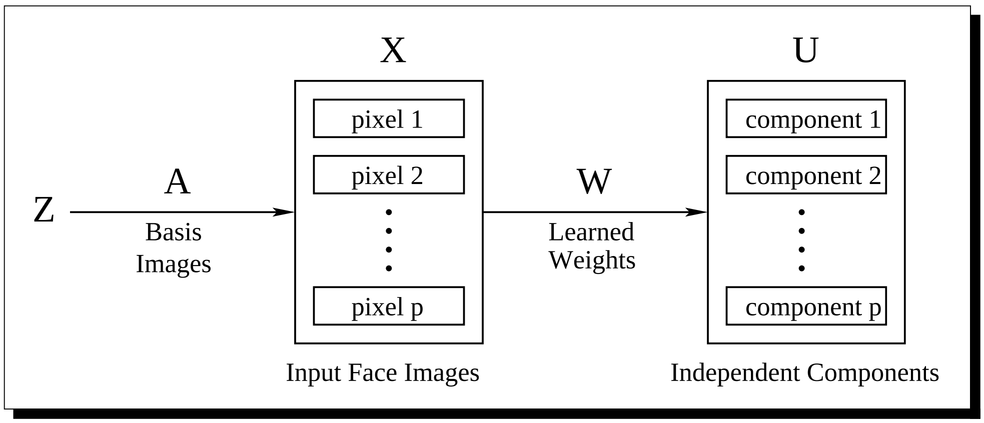


Figure 5.9: Finding statistically independent components.

This architecture was used in [16] to find image filters that produced statistically independent outputs from natural scenes. The eight basis images shown in the bottom row of 5.8 show more global properties than the basis images produced in architecture I (middle row). In this study, ICA is performed on the PCA components rather than directly on the input images to reduce the dimensionality as in [9, 10].

5.3.3 FA

Factor analysis [139] is a statistical multivariate analysis technique similar to PCA for explaining the correlations among observed variables in terms of a smaller number of unobservable variables. The crucial difference between FA and PCA is that FA is concerned with the common variance accounted for by linear factors excluding the noise due to unique variance. PCA, on the other hand, finds the basis vectors that optimally explain the total variance; the unique variance is not computed or accounted for separately.

In the general FA model, a p-dimensional, mean centered observed vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ is composed of a vector $\mathbf{z} = (z_1, \ldots, z_q)^T$ of q latent variables (factors), and the vector $\mathbf{n} = (n_1, \ldots, n_p)^T$ of p independent disturbance variables [11]:

$$\mathbf{x} = \mathbf{\Lambda z} + \mathbf{n} \tag{5.2}$$

where $\mathbf{\Lambda}$ is called the *factor loading matrix*, whose element λ_{ij} determines the importance of factor z_j to x_i. Since all factors are unobservable, the factor loadings provide the only means of 'labeling' each factor. The disturbance term $\mathbf{n}$ accounts for independent noise in each element of $\mathbf{x}$.

In this model, it is assumed that the underlying distribution of $\mathbf{z}$ is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{n}$ follows $\mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi} = diag(\psi_1, \ldots, \psi_p)$. Since it is assumed that n_i's are uncorrelated to each other, x_i's are conditionally uncorrelated given $\mathbf{z}$. The covariance matrix of $\mathbf{x}$ is computed as:

$$cov(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$$

Therefore, the variance of a variable is split into two parts:

$$var(x_i) = \sigma_i^2 = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i \tag{5.3}$$

The first term, the sum of squared factor loadings across factors, is called the common variance or *communality*, which is the variance accounted for by the factors. The second term, ψ_i is called *specific* or *unique* variance of x_i, which determines how much of the variability in each x_i is not attributable to the common factors. It is the variance due to n_i, not shared by other variables $x_j (j \neq i)$. The goal of FA model is to determine the $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ that best explain the covariance structure of $\mathbf{x}$ [11, 55].

In this work, we tested an EM algorithm for maximum likelihood estimation of FA proposed in [128] and reviewed in [55] and [129]. Let $\mathbf{X}$ be a data matrix whose rows are m observed sample vectors. Then, the expected log-likelihood for FA is

$$L(\mathbf{\Lambda}, \mathbf{\Psi}) \quad = \quad \log \left[\prod_{i=1}^{m} (2\pi)^{-p/2} |\mathbf{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right\} \right]$$

$$= -\frac{mp}{2}\log(2\pi) - \frac{m}{2}\log|\mathbf{\Sigma}| - \frac{m}{2}trace(\mathbf{C_x}\mathbf{\Sigma}^{-1})$$

where $\mathbf{C_x} = \mathbf{X}^T\mathbf{X}/m$. The EM algorithm maximizes L by going through two steps iteratively. In the *E-step*, $E[\mathbf{z}|\mathbf{x}_i]$ and $E[\mathbf{z}\mathbf{z}^T|\mathbf{x}_i]$, $i = 1, \ldots, m$, are computed by:

$$\begin{aligned}
E[\mathbf{z}|\mathbf{x}_i] &= \mathbf{W}\mathbf{x}_i \\
E[\mathbf{z}\mathbf{z}^T|\mathbf{x}_i] &= \mathbf{I} - \mathbf{W}\mathbf{\Lambda} + \mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T
\end{aligned}$$

where $\mathbf{W} = \mathbf{\Lambda}^T(\mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1}$. Then, in the *M-step*, new estimates for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are computed as follows:

$$\begin{aligned}
\mathbf{\Lambda}^{new} &= \left(\sum_{i=1}^{m}\mathbf{x}_i E[\mathbf{z}|\mathbf{x}_i]^T\right)\left(\sum_{j=1}^{m}E[\mathbf{z}\mathbf{z}^T|\mathbf{x}_j]\right)^{-1} \\
\mathbf{\Psi}^{new} &= \frac{1}{m}diag\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^T - \mathbf{\Lambda}^{new}E[\mathbf{z}|\mathbf{x}_i]\mathbf{x}_i^T\right)
\end{aligned}$$

(For the detailed derivation of the formula described in [55], see Appendix A.).

5.4 Summary

In this chapter, we proposed possible computational algorithms to implement category and exemplar pattern activation subsystems. Categorization is modeled as unsupervised clustering during training followed by nearest neighbor match during testing while exemplar recognition is modeled as subspace extraction during training and subspace projection followed by nearest neighbor match during testing. It builds a local linear model rather than a single, global model, which would be more suitable for expert object recognition paradigm where images of different objects are less likely to come from a single distribution.

We described three unsupervised clustering algorithms: K-Means, Gaussian Mixtures with traditional EM, and Gaussian Mixtures with probabilistically weighted PCA. K-Means is fast, simple, robust, and easy to apply to high-dimensional data, but implicitly models every data cluster with a symmetric Gaussian distribution. On the other hand, Gaussian mixtures with traditional EM can fit arbitrary asymmetric Gaussian distributions to model the data set. In practice, however, it is difficult to apply traditional EM to high-dimensional data because of numerical problems with inverting near singular matrices. This problem can be avoided to some degree by decomposing the probability into two parts corresponding to the principal subspace probability and the probability in the orthogonal complementary subspace. We illustrated how differently the three clustering algorithms work using a 2D synthetic data set.

Kosslyn suggests that objects are stored as "compressed images" in the visual memory and used for exemplar recognition. A possible way to implement "compressed image" representation is to map the data onto a lower dimensional space found by subspace projection algorithms. We described three different algorithms in this chapter; PCA is the most popular and widely used technique for data analysis, which decorrelates the input data using second-order statistics. ICA goes a step further to minimize higher-order dependencies in the input data in addition to linear decorrelation. By transposing the input data matrix, ICA can be applied to object recognition in two different ways. Lastly, FA is a similar technique to PCA except that it separates out the noise due to the unique variance in each dimension.

In the next chapter, we put all pieces together and present experimental results from running the complete expert object recognition system with multi-class data sets.

Chapter 6

An Expert Object Recognition System

We have now described the main components of an object recognition system based on a biological model of visual object recognition. We have presented how the role of the attention window is implemented as a patch extraction process, how to generate biologically plausible features from images, and how pattern matching mechanism can be implemented using unsupervised clustering and local subspace projection. In this chapter, we link these pieces together to make a complete end-to-end object recognition system and discuss how the system works by running the system on four multi-class data sets: one of them is a simple synthetic data set containing a set of 2D points, while the other three data sets contain real images.

Although the system is composed of distinct components, they are linked by feed-forward connections and therefore any change to one component may affect the system's overall behavior. In the experiments presented in this chapter, however, we focus on the pattern matching mechanism. Following Kosslyn's biological model, the system implements pattern matching using two separate, hierarchically connected components: classification and exemplar matching. The classification and exemplar matching modules can be considered as standard object recognition systems by themselves. The experimental goal is to analyze the effectiveness of designing the system using combinations of classification and exemplar matching components. For that, the system's overall performance is compared to the performance of its components alone.

The experiments are performed in the context of expert object recognition. As described in Section 3.2, expert object recognition is viewpoint dependent. Here, the system is trained for given classes of objects viewed from fixed viewpoints, and tested with familiar objects viewed from the same viewpoints as the training objects. Another characteristic of expert object recognition is multiple-level categorization. The objects in a data set of aerial images are defined using a hierarchy of categories so that we can test the system for recognizing objects in different levels of categories.

We made independent studies on subspace projection algorithms as presented in the next chapter. The results indicate that ICA sometimes outperforms PCA, however, the performance of ICA varies depending on the architecture and the task. Compared to ICA, PCA shows more stable performance regardless of the nature of the task. Moreover, PCA is robust, runs faster than ICA, and is considered as a standard subspace projection technique. Therefore, the current system implements exemplar matching by PCA. For categorization, all three clustering algorithms were tested on the synthetic data set, but traditional EM was not applied for the other two data sets because of the dimensionality limitation.

6.1 Experiments

6.1.1 Performance on 2D Synthetic Data Set

The synthetic data set used in this experiment contains 1,400 2D points randomly sampled from two Gaussian distributions as described in Section 5.2.1. Each Gaussian distribution generates 700 data points, and the set is divided into a training set of 800 points and a test set of 600 points (Figure 6.1). Data values are assumed to be the extracted features so categorization is directly applied to the data set.

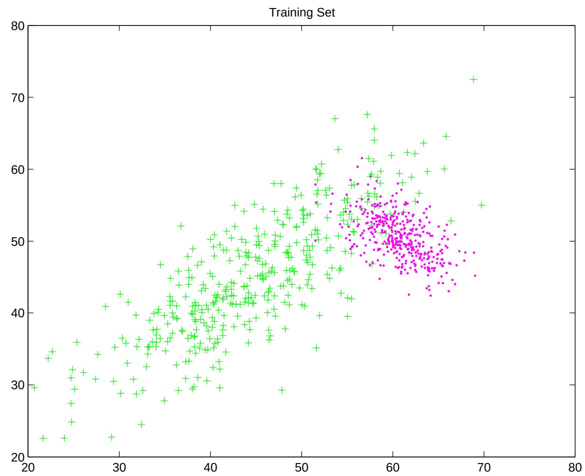


Figure 6.1: The 2D synthetic training (left) and test sets (right). Point patterns are different according to the underlying Gaussian distribution.

Table 6.1 shows the recognition rates for three versions of the proposed system, each using K-Means, traditional EM, and PWPCA clustering. The number of clusters is two and data points are projected to the first principal axis to find the closest match. All clustering algorithms ran for 10 iterations. Table 6.1 also shows the results of PCA analysis without clustering, and for clustering without PCA, where points are assigned to the label of the Gaussian distribution which produces the majority of the points in the cluster. Table 6.2 shows the confidence in the hypothesis that the version shown in the left-most column is more accurate than the versions shown in the same row, as measured by McNemar's significance test [158] for paired binomial values.

As can be seen from the tables, best performance was obtained using traditional EM without PCA, which performs significantly better than all three versions of the system. This result can be explained from the structure of the data set. The data set is a mixture of two Gaussian distributions, and EM finds nearly perfect clusters as shown in Figure 5.3. A hard assignment is applied after the final iteration, so some points from the elongated Gaussian distribution in the crossing area are assigned to the wrong cluster. When PCA is performed in the cluster, these points cause incorrect matches for some points that would otherwise be correctly classified by the clustering algorithm. As a result, the performance of the overall system drops. While traditional EM uses exact probability computation, PWPCA clustering approximates the probability as shown in equation (5.1). However, when the data dimension is reduced from two to one as in this experiment, there is no error caused by the approximation. Although PWPCA clustering seems to produce slightly less accurate clustering result than the traditional EM, the difference is not statistically significant as shown in Table 6.2.

	Overall (KM)	Overall (EM)	Overall (PWPCA)	Global PCA	Clustering (KM)	Clustering (EM)	Clustering (PWPCA)
Recog. rates	86.3%	92.3%	92.7%	86.8%	82.8%	95.3%	94.2%

Table 6.1: Recognition rates of the system using K-Means, traditional EM, and PW-PCA clustering followed by PCA, along with PCA without clustering and clustering without PCA.

	Overall (KM)	Overall (EM)	Overall (PWPCA)	Global PCA	Clustering (KM)	Clustering (EM)	Clustering (PWPCA)
Overall (KM)	-	-	-	-	98%	-	-
Overall (EM)	99%	-	-	99%	99%	-	-
Overall (PWPCA)	99%	56%	-	99%	99%	-	-
Global PCA	58%	-	-	-	98%	-	-
Clustering (EM)	99%	99%	99%	99%	99%	-	85%
Clustering (PWPCA)	99%	94%	92%	99%	99%	-	-

Table 6.2: Confidence evaluated by the McNemar's test on the hypothesis that the version of the system shown in the left-most column is more accurate than the versions shown in the same row.

When K-Means is used, the results are quite different. The overall performance of the system is significantly better than with clustering alone, but it does not improve the performance over global PCA. This result can be expected again from the distribution of the data points; the principal axis of the data set as a whole is approximately 45 degrees, so most of the errors are from points in the top half region. As we can see in Figure 5.3, K-Means was not able to separate data points in that region, therefore it performs similarly to global PCA. On the other hand, traditional EM and PWPCA clustering can fit the underlying Gaussian distributions very closely (Figure 5.3). As a result, the recognition rates using the two algorithms are significantly better than global PCA and the version using K-Means.

The experiment discussed in this section shows how the system would behave on data sets that are perfect mixtures of Gaussians. The EM and PWPCA clustering algorithms fit the underlying distributions, and therefore produce near optimal classification results. In this case, combining classification and exemplar matching does not perform better than the classification alone, however, it still improves the performance over global PCA. If the underlying Gaussian distributions are asymmetric, using K-Means for clustering does not help the system at all. However, if we increase the number of clusters K-Means may perform better, since an asymmetric Gaussian can be approximated by multiple symmetric Gaussian distributions. The next two experiments are performed on real data sets, where the underlying distributions are unknown.

6.1.2 Performance on Real Data Sets

6.1.2.1 Data Sets

Cat and Dog Database: This data set contains 200 64×64 black and white images of cat faces and dog faces – 100 cats and 100 dogs. No subjects are repeated in the database. The images were collected from the web, and exhibit a wide range of backgrounds and illumination conditions. The images were registered by hand to within a similarity transform – translation, rotation, and scale – by roughly aligning the eyes of the faces in all 200 images. This data set is challenging because although there are differences between cats and dogs, there are also many variations among instances of each animal class. Figure 6.2 shows some examples of images from the data set.

Since each image in the data set contains one subject and is approximately registered to the other images, no focus of attention process is needed. Images are directly presented to the feature generation system, where the average complex edge magnitudes and Hough space representations are computed.

Figure 6.2: Sample images from the Cat and Dog data set. The top row is all cats and the bottom row is all dogs.

Ft. Hood Imagery Data Set: The Ft. Hood Data set is a moderately large data set containing 8-bit gray-scale images of size roughly 7700×7700 collected around the Ft. Hood, TX area. In the data set, there are seven vertical-view images, which were taken with the camera pointing down almost strictly vertically[1]. In this study, we use two vertical-view images containing several different types of objects. Figure 6.3 shows a small image cut from an original large size image.

As we can see in the figure, it contains different styles of buildings, parking lots, roads, and sidewalks. To generate patch images containing interesting objects, the

[1]The official description of the data set and the images can be obtained at http://www.mbvlab.wpafb.af.mil/public/sdms/datasets/fthood/.

Figure 6.3: A sample Ft. Hood image of size 1927 × 1922.

Ft. Hood images are fed into the patch extraction system described in Section 4.1. The extracted patch images are hand-labeled according to the object within them. The labels are defined in multiple levels. For example, at the highest level, objects can be differentiated as buildings or fields. And those objects defined as a field can be separated as natural grounds or parking lots. Parking lots can be further divided into paved and unpaved parking lots.

Figure 6.4 shows patch images of six different objects. For the experiments performed in this study, we created two data sets from Ft. Hood images, both containing four different types of objects. The first data set, referred to as $FH1$, contains 600 images of an industrial building style, paved parking lot, natural ground, and sidewalks. The second data set, called $FH2$, contains 600 images of two styles of industrial building and paved/unpaved parking lots. Therefore, at a higher level, the FH2 data set contains two classes of objects – industrial buildings and parking lots – while four classes are still defined for the corresponding level in FH1 data set. The number of images of each object is equal in both data sets.

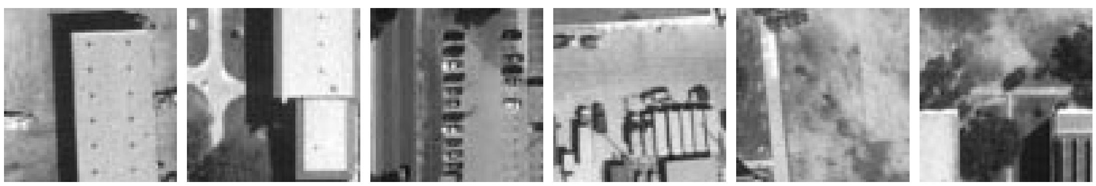


Figure 6.4: Example patch images extracted from the Ft. Hood data set. The two leftmost images contain different styles of industrial building, the two middle images contain paved and unpaved parking lot, and the remaining two images show natural ground and sidewalk.

6.1.2.2 Feature Extraction

To compute the average complex edge magnitudes at multiple scales, images are successively scaled down by a factor of two until they get to 8×8 pixels. Then, a bank of Gabor filters are applied and the resulting even and odd filter responses are combined to produce the Gabor energy output. (For a more detailed description, see Chapter 4.) The average edge magnitude is computed by combining six Gabor energies computed every 15 degrees (Figure 4.6 and Figure 4.7). This process is applied at every scale from 32×32 pixels to 8×8 pixels, generating a total of 1,344 feature values.

The Hough space representation described in Section 4.3.2 is computed by quantizing the parameters r and w at every five pixels and 15 degrees, respectively. Since the maximum perpendicular distance from the origin to a line in a 64×64 image is $64\sqrt{2}$, r is divided into 18 bins and w is divided into 24 bins. The Hough transform is applied to the edge images computed at every 45 degrees, and combined into one Hough space histogram without thresholding. The resulting non-zero votes in the 432 bins are the final feature values to be used in pattern matching.

6.1.2.3 Recognition Results: Cat and Dog data set

For the Cat and Dog data set, the system was tested 25 times, each time training on 160 randomly selected images and testing on the remaining 40 images. The system maps test images onto images in memory (training images) and assigns labels to objects by replicating the semantic labels of the retrieved training image. A test image was recognized correctly if the retrieved training image was of the same species as the test image. With 40 test images per trial and 25 trials, each version of the system was evaluated on a total of 1,000 matches.

Table 6.3 shows the recognition rates on average edge magnitude features for two versions of the system, one using K-Means and the other using PWPCA clustering, along with the results for each component alone – PCA without clustering and clustering without PCA. Traditional EM can not run on the real data sets because of the dimensionality problem. For clustering without PCA, images are assigned to the species that is dominant in their cluster. This is only for evaluation; the system trains and runs without supervision. Since there are two object types in the domain, the number of clusters was two except for the third and last column, where K-Means was run with five clusters. In all cases, the number of PCA subspace dimension was 10. Table 6.4 shows the confidence level for pairs of techniques.

The results in Table 6.3 show that, using PWPCA clustering, the system as a whole outperforms any of its components. The performance of clustering alone is significantly worse than the overall system, which indicates that the classes are not drawn from cleanly separated Gaussian distributions. The clusters found by the categorization system contains mixed set of objects, and the following exemplar match refines the system's performance. With K-Means, the system still outperforms categorization alone, but makes no difference over exemplar matching without clustering. However, the performance of the system using K-Means improves if the number of clusters is increased to five, even though there are only two object classes in the data set, presumably because asymmetric Gaussian distributions can be approximated using multiple symmetric distributions.

	Overall (KM)	Overall (PWPCA)	Overall (KM:K=5)	Global PCA	Clustering (KM)	Clustering (PWPCA)	Clustering (KM:K=5)
Recog. rates	77.3%	80.5%	79.1%	77.5%	64.8.8%	60.7%	71.6%

Table 6.3: Recognition rates of the system for the Cat and Dog data set on average complex edge magnitude using K-Means, PWPCA, and K-Means with five clusters followed by PCA, along with PCA without clustering and clustering without PCA. Except for the third and last column, the number of clusters was two, and the subspace dimension was 10 for all cases.

	Overall (KM)	Overall (PWPCA)	Overall (KM:K=5)	Global PCA	Clustering (KM)	Clustering (PWPCA)	Clustering (KM:K=5)
Overall (KM)	-	-	-	-	99%	99%	99%
Overall (PWPCA)	98%	-	83%	99%	99%	99%	99%
Overall (KM:K=5)	91%	-	-	86%	99%	99%	99%
Global PCA	53%	-	-	-	99%	99%	99%
Clustering (KM)	-	-	-	-	-	97%	-
Clustering (KM:K=5)	-	-	-	-	99%	99%	-

Table 6.4: Confidence evaluated by the McNemar's test on the hypothesis that the version of the system shown in the left-most column is more accurate than the versions shown in the same row.

Figure 6.5 through Figure 6.8 show the recognition rates on both average edge magnitude features and Hough space features for a range of number of clusters K and subspace dimension q. In all graphs presented here, comparisons between the system's performance and the categorization component alone are omitted since the system always significantly outperforms the categorization component. For average edge magnitude features, it can be seen that the performance of the K-Means version of the system improves as K increases. For Hough features, the performance is still better when K is larger than two.

Another thing that can be noted from the graphs is that, for a range of small q's, both versions of the system performs better than global PCA. After the performance of the system crosses with the performance line for global PCA, the PWPCA clustering version of the system mostly performs worse than global PCA. The K-Means version of the system, on the other hand, keeps doing a better or comparable job for larger q values. In general, the system works more effectively when more information in the data set is not kept; that is, when the data compression rate is high.

For the results for categorization alone, an interesting observation can be found from Table 6.3. It shows that K-Means does a better job than PWPCA clustering for assigning images to clusters according to their symbolic labels. However, the overall performance of the system using K-Means is lower than using PWPCA clustering, which indicates that the categories found by PWPCA clustering may not match the symbolic object labels. As can be seen in Figure 6.2, there are cat-like dogs with pointy ears. Images of these cat-like dogs may have similar underlying distribution to cat images in feature space and PWPCA clustering might group these images with the cats, producing mixed clusters in terms of symbolic labels. The exemplar recognition, however, matches cat-like dogs to other cat-like dogs, as a result, it produces a correct classification.

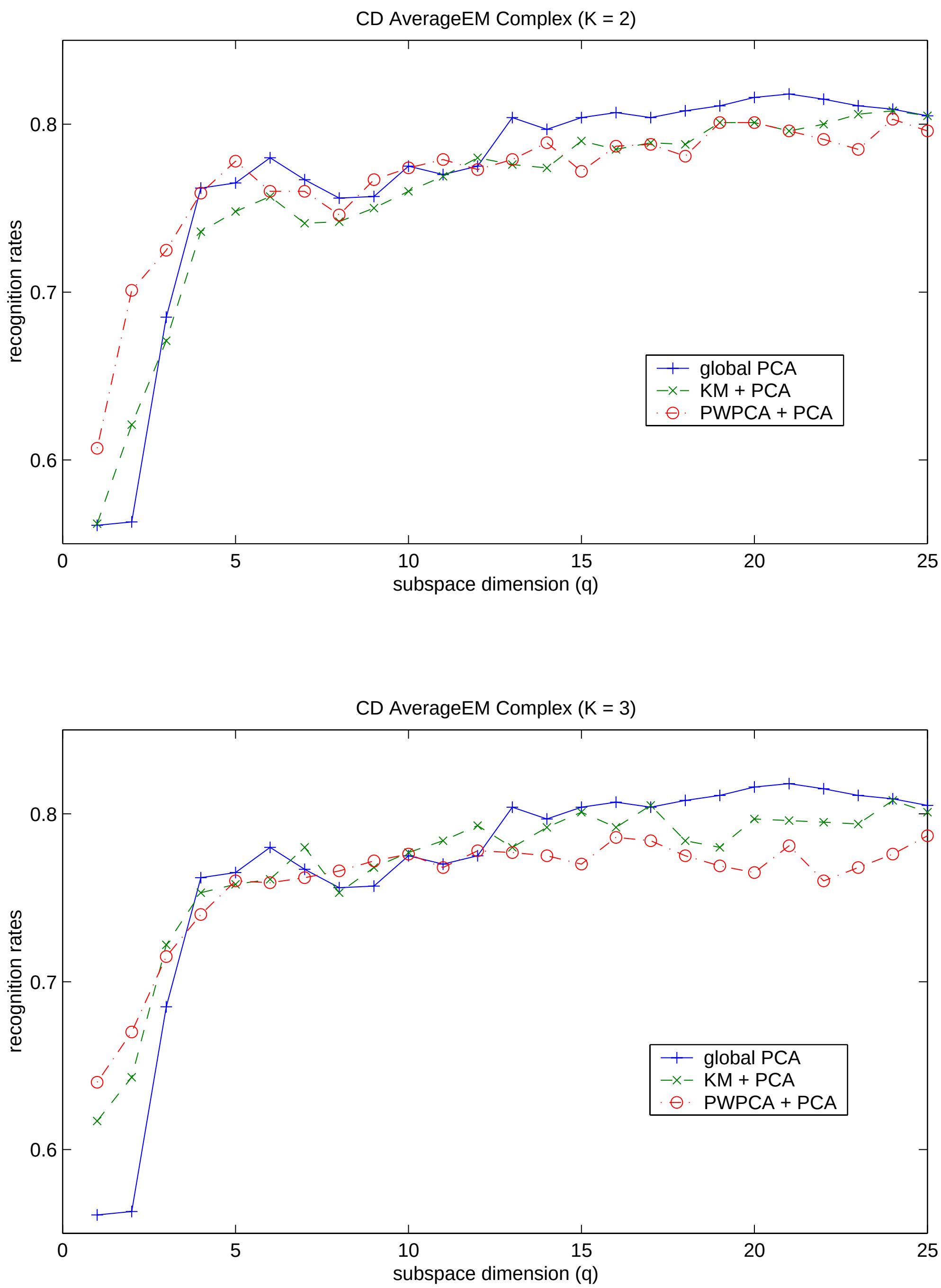
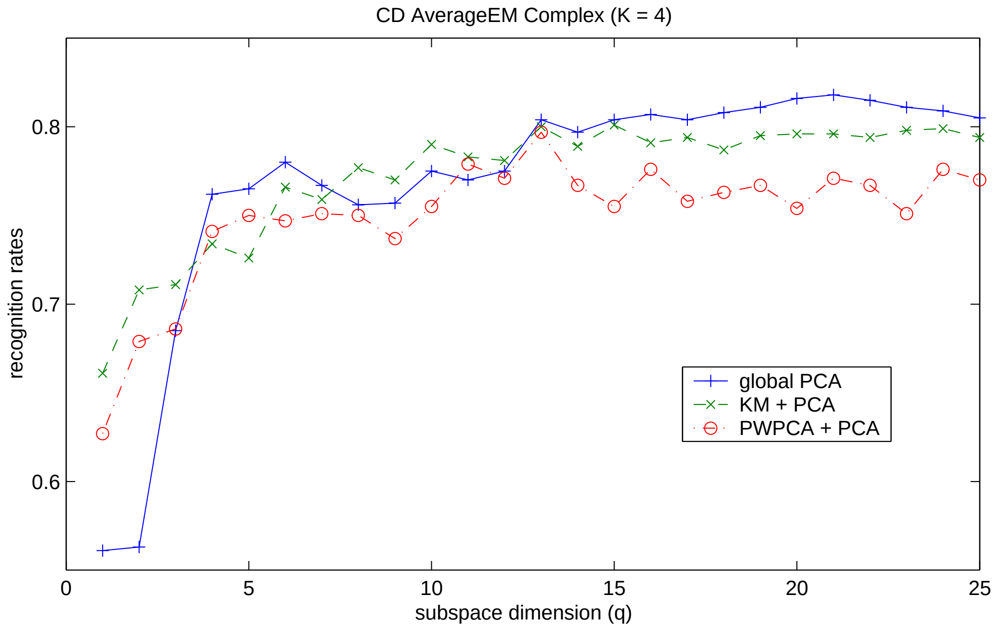


Figure 6.5: Recognition rates for the two versions of the system and global PCA on average complex edge magnitude of the Cat and Dog data set for $K = 2$ (top) and $K = 3$ (bottom). The subspace dimension q varies from 1 to 25.



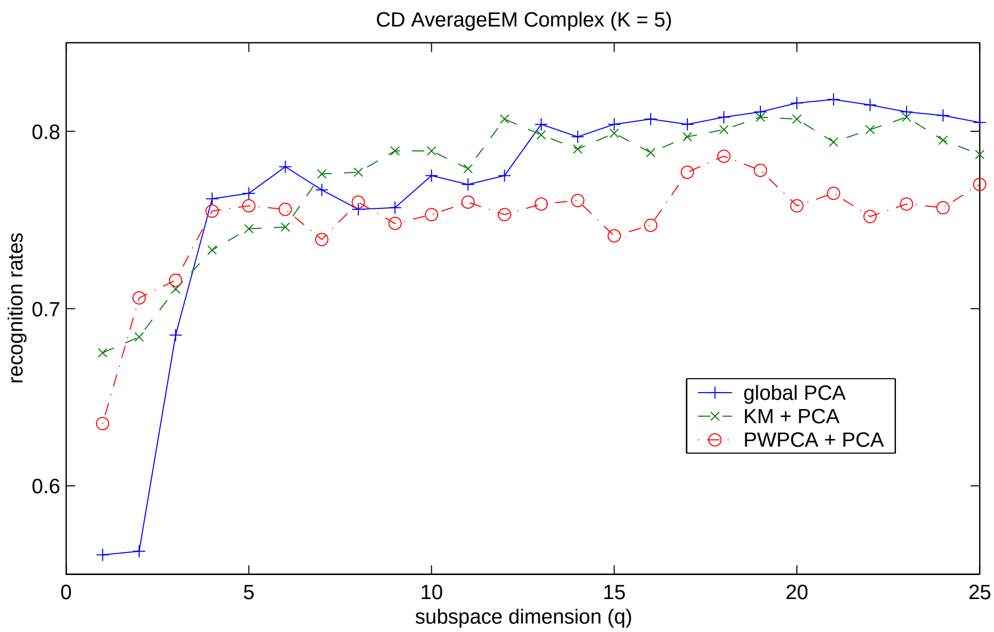
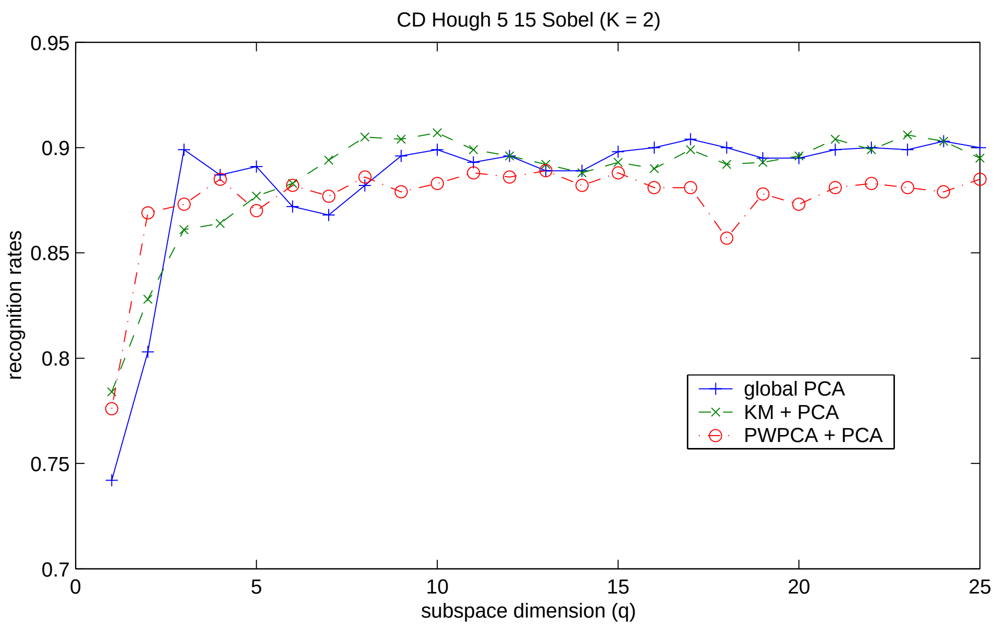


Figure 6.6: Recognition rates for the two versions of the system and global PCA on average complex edge magnitude of the Cat and Dog data set for $K = 4$ (top) and $K = 5$ (bottom). The subspace dimension q varies from 1 to 25.



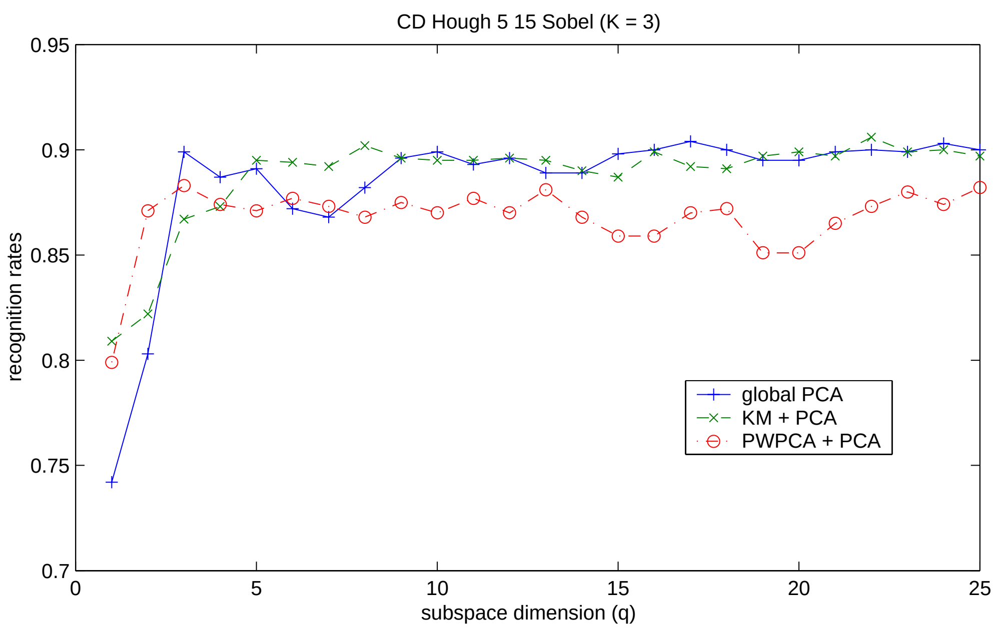


Figure 6.7: Recognition rates for the two versions of the system and global PCA on Hough space features of the Cat and Dog data set for $K = 2$ (top) and $K = 3$ (bottom). The subspace dimension q varies from 1 to 25.

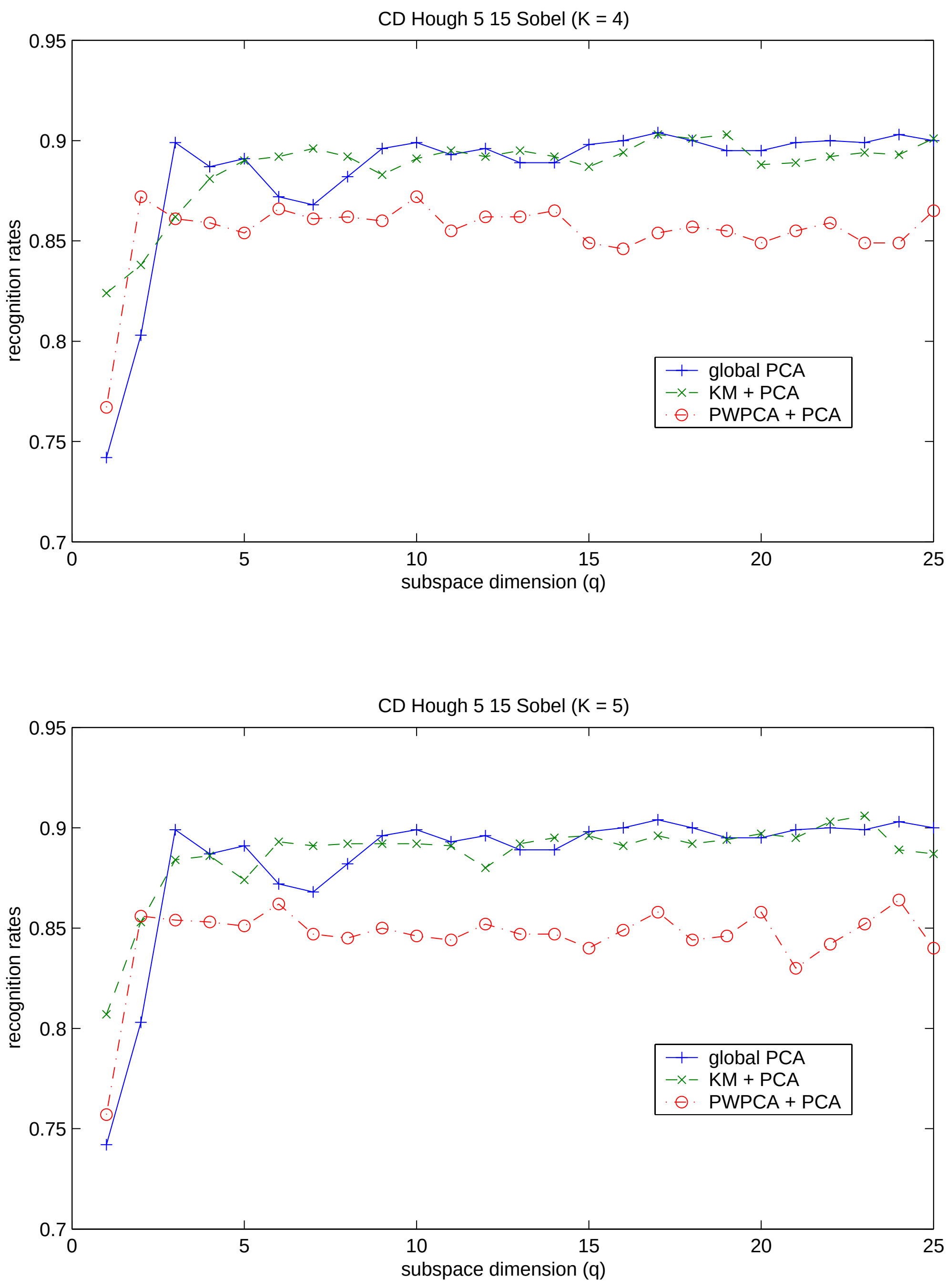


Figure 6.8: Recognition rates for the two versions of the system and global PCA on Hough space features of the Cat and Dog data set for $K = 4$ (top) and $K = 5$ (bottom). The subspace dimension q varies from 1 to 25.

6.1.2.4 Recognition Results: Ft. Hood data sets

When we tested the system on Ft. Hood data sets, we faced the numerical problem described in Appendix B for probability computation in PWPCA clustering. The range of exponent in the outer-subspace probability term is very sensitive to the average trailing variance σ which varies across clusters in every iteration. So, as described in Appendix B, we assume that the average trailing variance are equal for every cluster in each iteration and set σ as the average of σ's of all clusters. From now on, the performance of the system with PWPCA clustering is obtained using this version. The effect of using different σ will be discussed later in this section.

The Ft. Hood data sets are divided into a training set containing 400 images and a test set containing 200 images. The patch images in the training and test sets were extracted from different parts of a vertical-view image. Therefore, there is no overlap between training and test sets. To test the system for the Ft. Hood data sets, we ran each version of the system 250 times for different number of clusters using 10 subspace dimensions, therefore evaluate the system on a total of 50,000 matches.

Table 6.5 and Table 6.6 show the average recognition rates for average complex edge magnitude features of the FH1 and FH2 data set, respectively. As we can see in Table 6.5, the results for the FH1 data set are similar to those for the Cat and Dog data set; using PWPCA clustering, the system as a whole outperforms any of its components. However, the results do not hold for the FH2 data set shown in Table 6.6. Although the two versions of the system with either clustering methods still outperform the categorization component alone and the version using K-Means is comparable to global PCA, the version with PWPCA clustering performs significantly worse than global PCA.

While seeking an explanation for the different results on the Ft. Hood data sets, we tried two things: (1) explore the parameter space for the subspace dimension and (2) compare PWPCA clustering with non-constant σ – the original implementation

	Global PCA	Overall (KM)	Overall (PWPCA)	Clustering (KM)	Clustering (PWPCA)
K=1	41.5%	-	-	-	-
K=2	-	42.37%	42.60%	29.51%	26.48%
K=3	-	39.33%	43.89%	36.61%	24.57%
K=4	-	42.05%	43.77%	39.35%	24.53%
K=5	-	42.69%	44.28%	39.73%	28.90%
K=6	-	42.90%	43.24%	40.60%	27.44%

Table 6.5: Recognition rates of the system for FH1 data set on average complex edge magnitude using PCA without clustering, K-Means and PWPCA clustering followed by PCA, and clustering without PCA for different number of clusters. The results are from a total of 250 runs with subspace dimension 10.

	Global PCA	Overall (KM)	Overall (PWPCA)	Clustering (KM)	Clustering (PWPCA)
K=1	62.5%	-	-	-	-
K=2	-	62.21%	57.52%	47.34%	26.05%
K=3	-	61.60%	57.36%	37.98%	24.20%
K=4	-	60.06%	56.30%	37.91%	24.12%
K=5	-	59.01%	54.53%	40.34%	28.86%

Table 6.6: Recognition rates of the system for FH2 data set on average complex edge magnitude using PCA without clustering, K-Means and PWPCA clustering followed by PCA, and clustering without PCA for different number of clusters. The results are from total of 250 runs with subspace dimension 10.

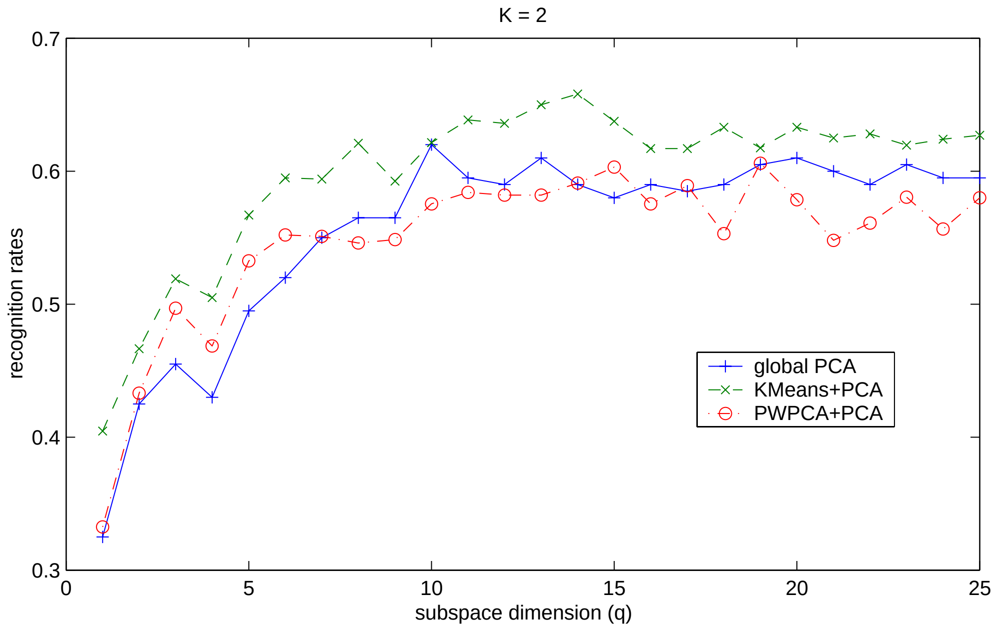
– and constant σ to see how the performance changes with the modification we made in the implementation to run the algorithm for Ft. Hood data sets.

Figure 6.9 through Figure 6.14 show recognition rates for the two versions of the system and global PCA on the Ft. Hood data sets across a range of subspace dimension q. The results were obtained by running each version of the system 10 times for every combination of K and q. As we have seen with the Cat and Dog data set, the graphs show that both versions of the system performs better than global PCA for a range of small q's. In most cases, the performance of the system using PWPCA clustering crosses with the performance line for global PCA more quickly than K-Means version, and from that point, mostly performs worse than global PCA. The K-Means version of the system, on the other hand, keeps doing a better job than global PCA even for larger q values if not comparable to it.

The results for using two different implementations of PWPCA clustering in the system are shown in Figure 6.15. In one implementation, the probability defined in equation (B.2) is computed using a separate σ for each cluster, while the other implementation uses a constant σ, the average value of σ's of all clusters, for the probability computation. Even though the entire training samples are used in PWPCA clustering, the intrinsic dimensionality of each cluster can differ depending on the weights of the samples. Since the in-subspace dimension is constant for all clusters, using a separate σ essentially discards different amounts of error for each cluster causing the clustering process to be unstable. That is why a constant, average σ was considered. For this comparison, the system was run on the Cat and Dog data set since both version of PWPCA clustering runs without any numerical problem for the data set. The figure shows that using a constant σ does not change the overall performance in general.

As described in Appendix B, σ has the role of balancing the influence of in-subspace distance and outer-subspace distance in the probability computation. For example, if σ is very small, that is most of the variances are kept within the principal subspace, then the probability would be very sensitive to changes in outer-subspace distance. Therefore, the importance of outer-subspace probability in equation (B.2) will vary depending on the magnitude of σ used.

We tested the system for several σ values by multiplying different weight values to the average σ. Figure 6.16 shows how the recognition rates changes as σ varies when 5, 10, 15, and 20 subspace dimensions are used. Although the graphs are not monotonic and change differently for each q, it seems that weight values smaller than one produce better results in general; the best performance is obtained using weights



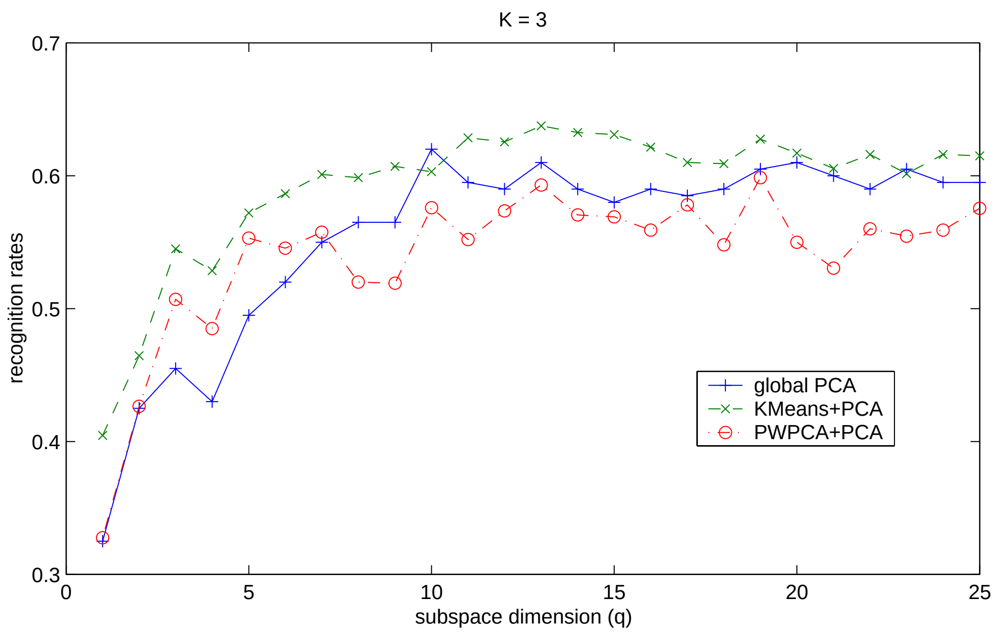


Figure 6.9: Recognition rates for the two versions of the system and global PCA on the average complex edge magnitude features of the FH2 data set for $K = 2$ (top) and $K = 3$ (bottom). The subspace dimension q varies from 1 to 25.

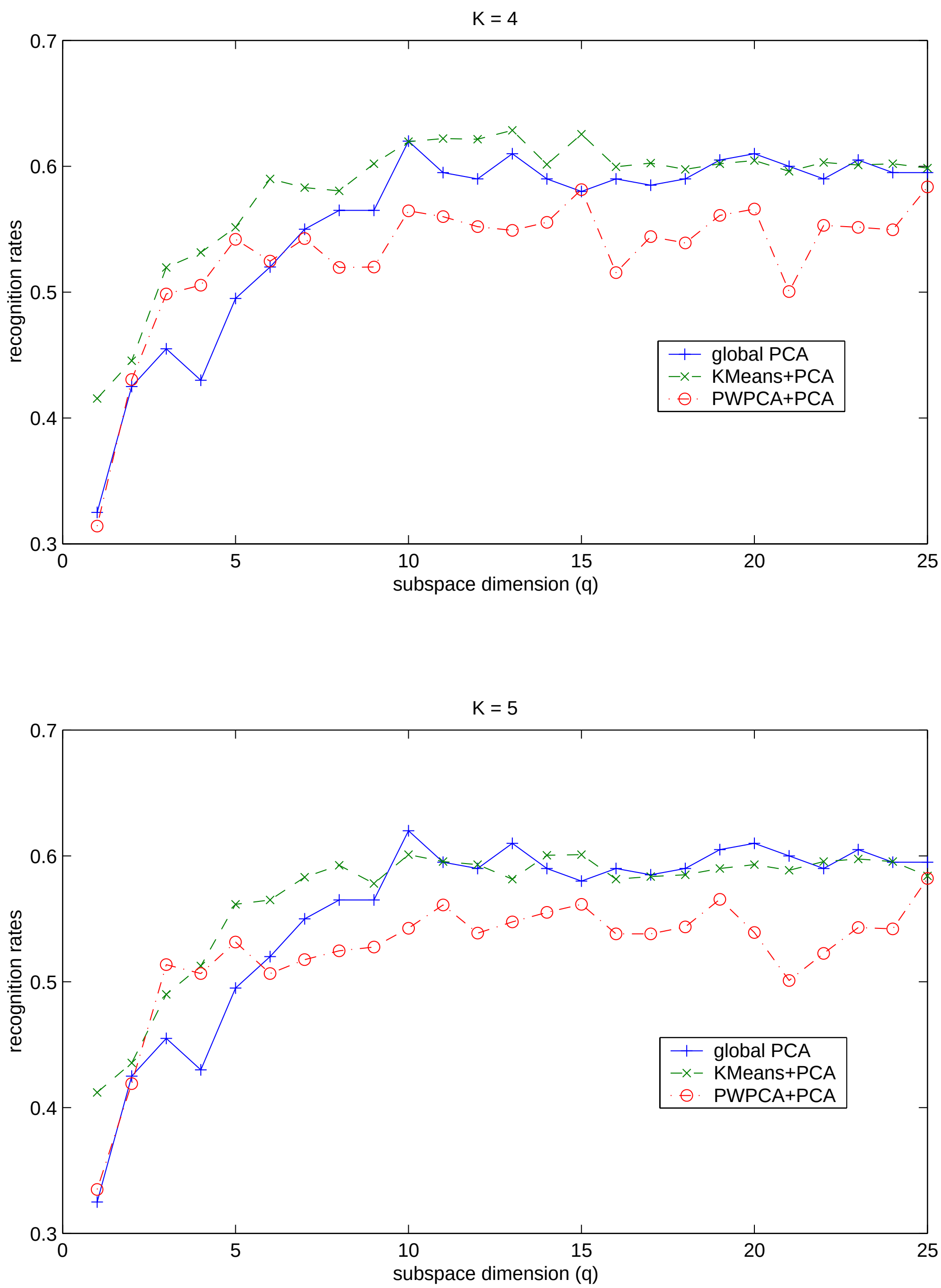


Figure 6.10: Recognition rates for the two versions of the system and global PCA on the average complex edge magnitude features of the FH2 data set for $K = 4$ (top) and $K = 5$ (bottom). The subspace dimension q varies from 1 to 25.

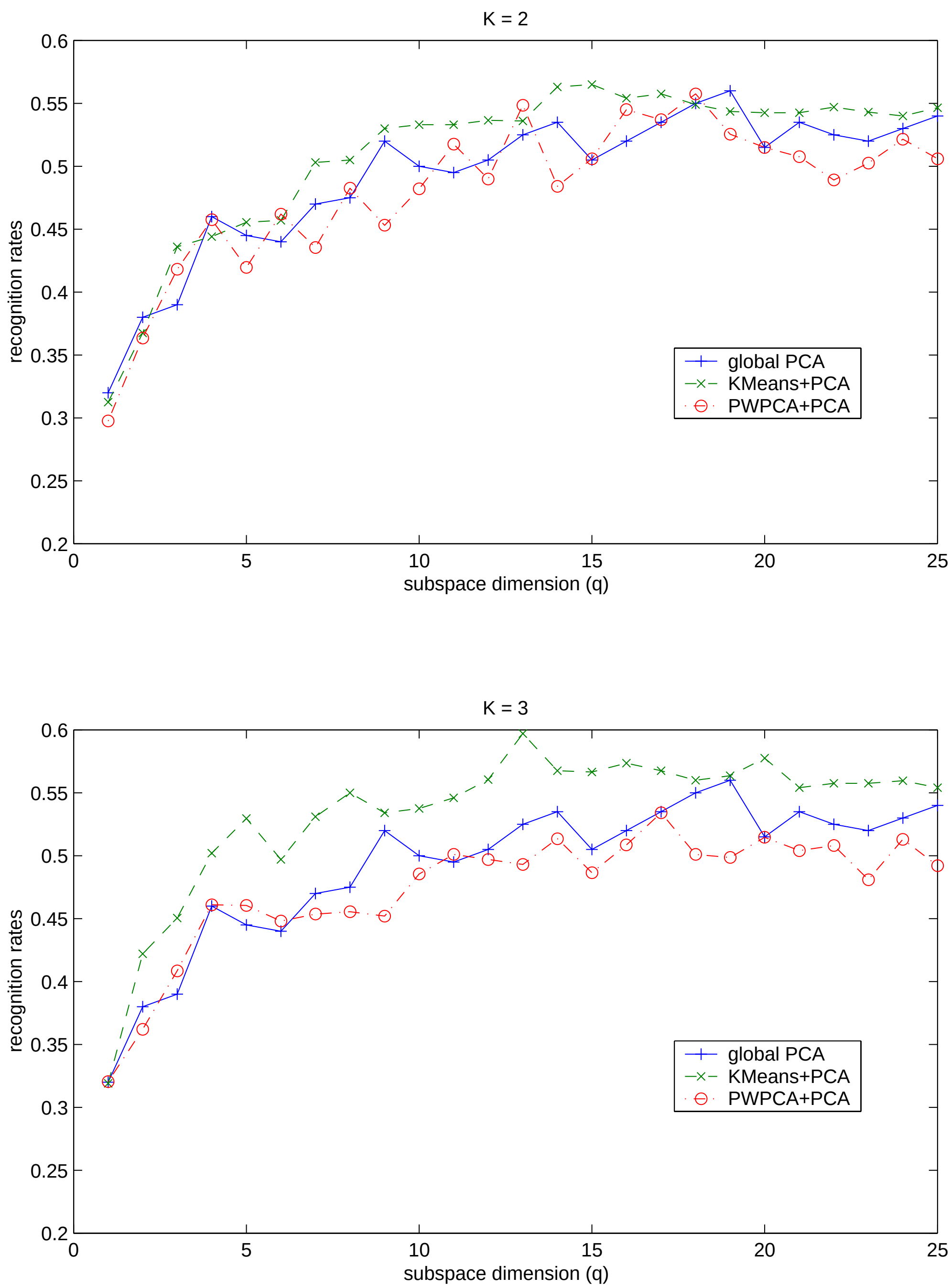


Figure 6.11: Recognition rates for the two versions of the system and global PCA on the Hough space features of the FH1 data set for $K = 2$ (top) and $K = 3$ (bottom). The subspace dimension q varies from 1 to 25.

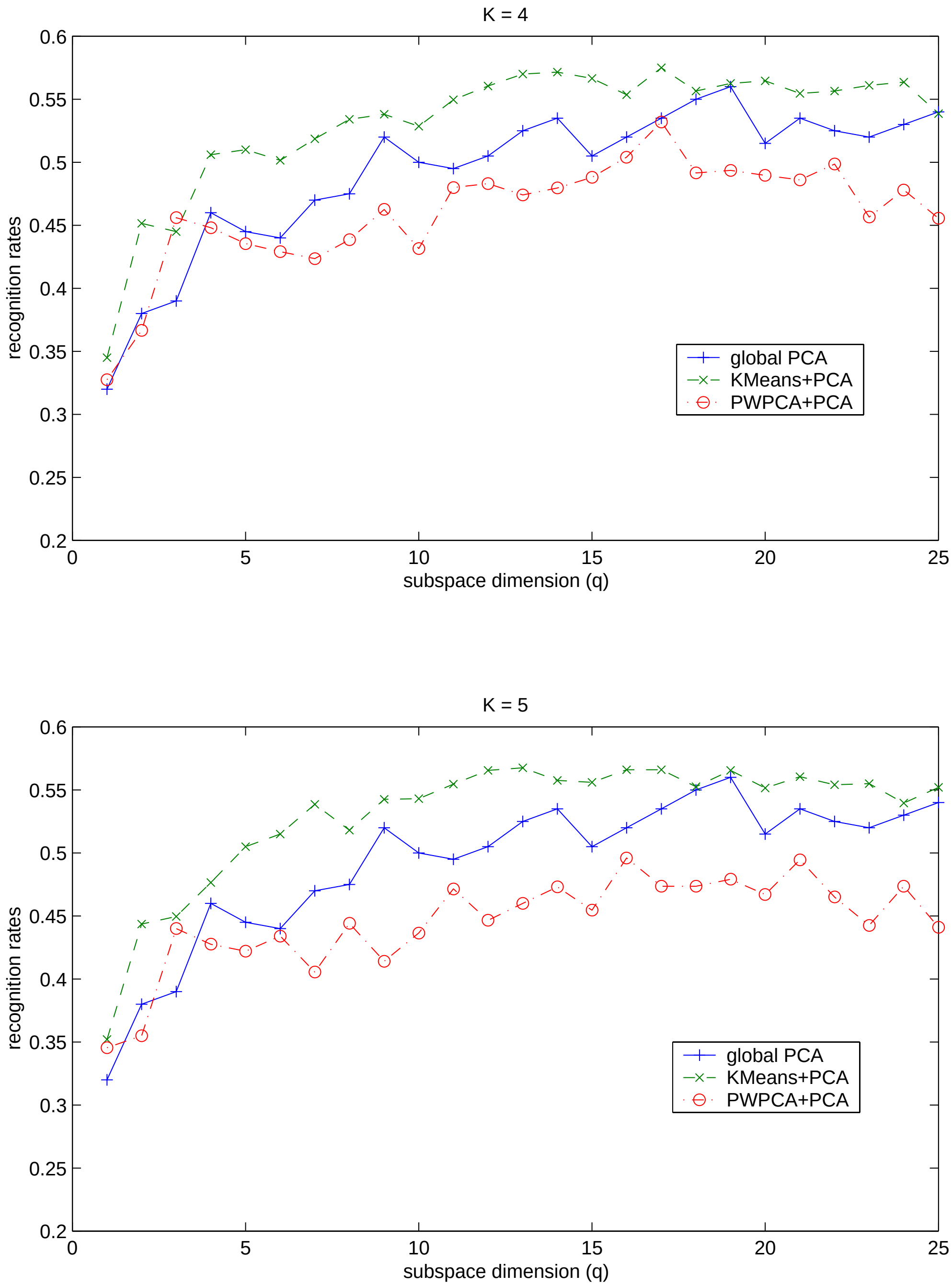


Figure 6.12: Recognition rates for the two versions of the system and global PCA on the Hough space features of the FH1 data set for $K = 4$ (top) and $K = 5$ (bottom). The subspace dimension q varies from 1 to 25.

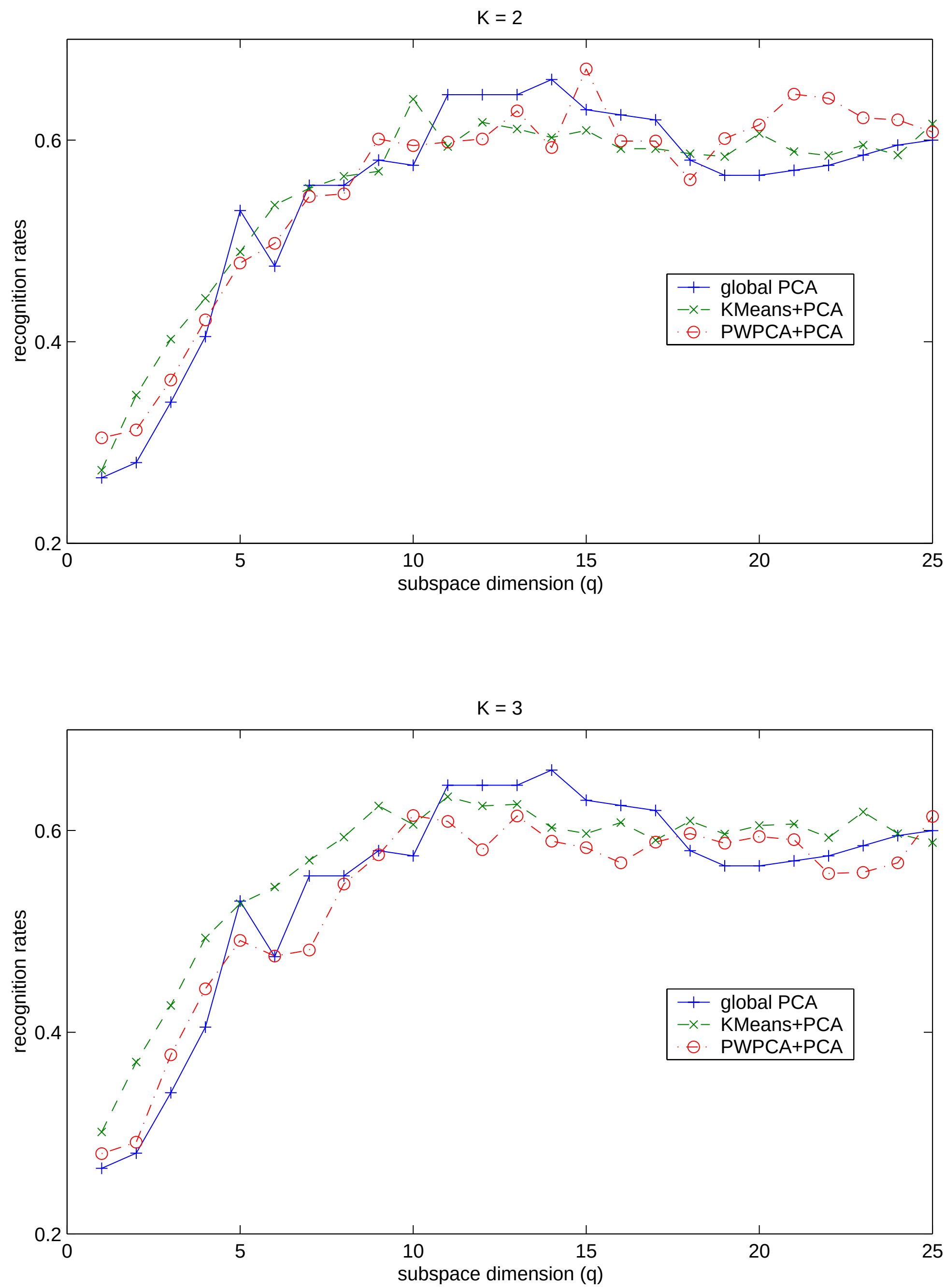
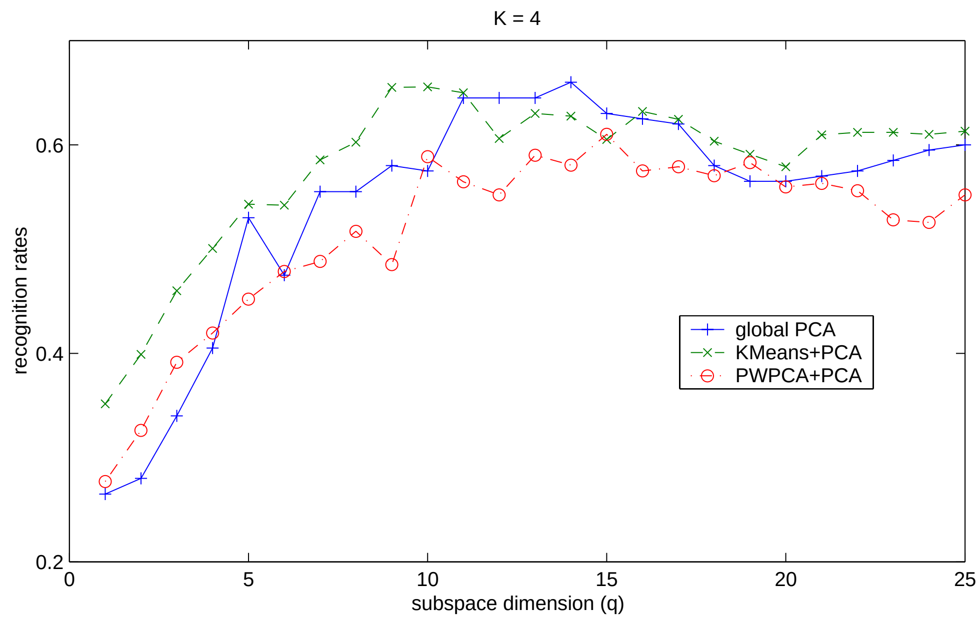


Figure 6.13: Recognition rates for the two versions of the system and global PCA on the Hough space features of the FH2 data set for $K = 2$ (top) and $K = 3$ (bottom). The subspace dimension q varies from 1 to 25.



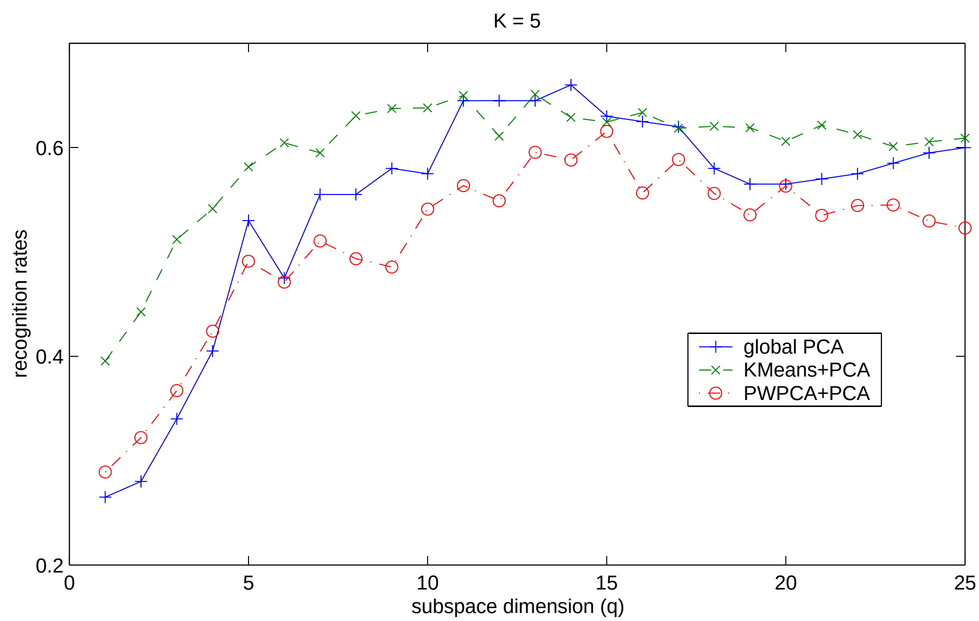


Figure 6.14: Recognition rates for the two versions of the system and global PCA on the Hough space features of the FH2 data set for $K = 4$ (top) and $K = 5$ (bottom). The subspace dimension q varies from 1 to 25.

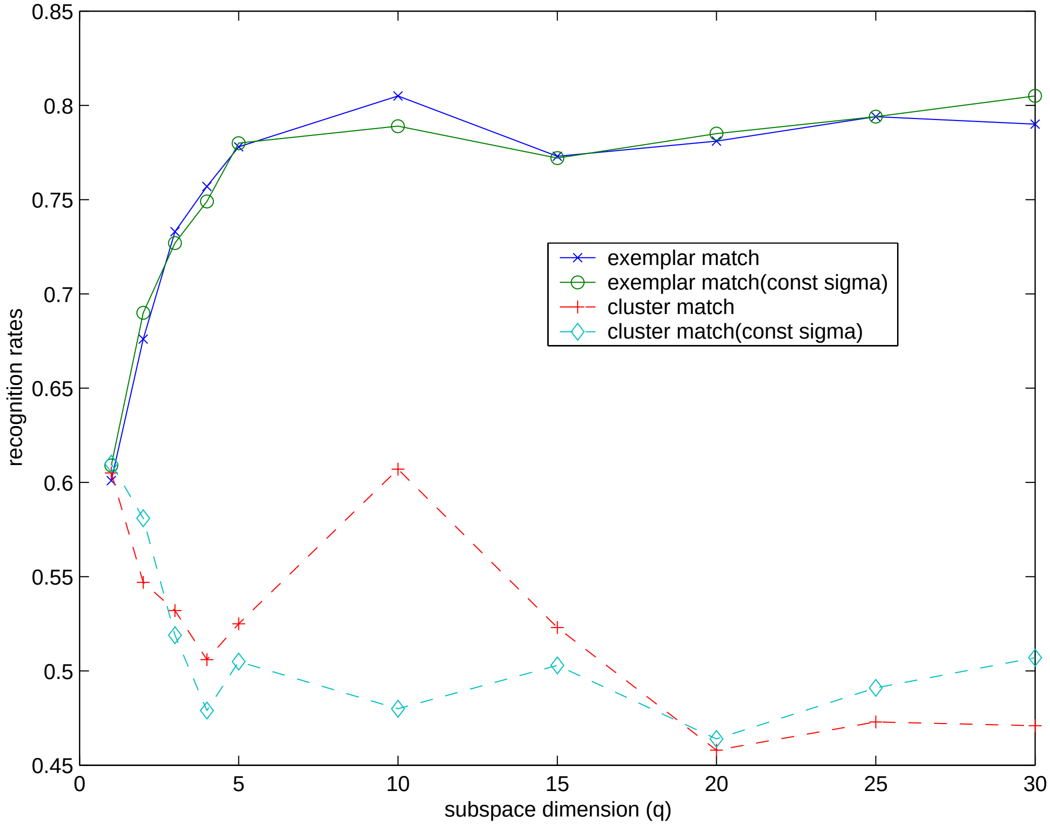


Figure 6.15: Recognition rates of the system on the Cat and Dog data set using two different implementations of PWPCA clustering for $K = 2$. Solid lines show the exemplar match results while dashed lines show the results for assigning dominant cluster labels.

less that one except for $q = 10$, in which case the performance for the weight of 0.75 is close to the best result. The results indicate that the system using PWPCA clustering can behave differently depending on how in-subspace and outer-subspace terms are balanced and the optimal balancing point may change as different subspace dimensions are used.

6.2 Summary

In this chapter, we presented experimental results of running the complete system on four multi-class data sets. The experiments focus on the pattern matching mechanism

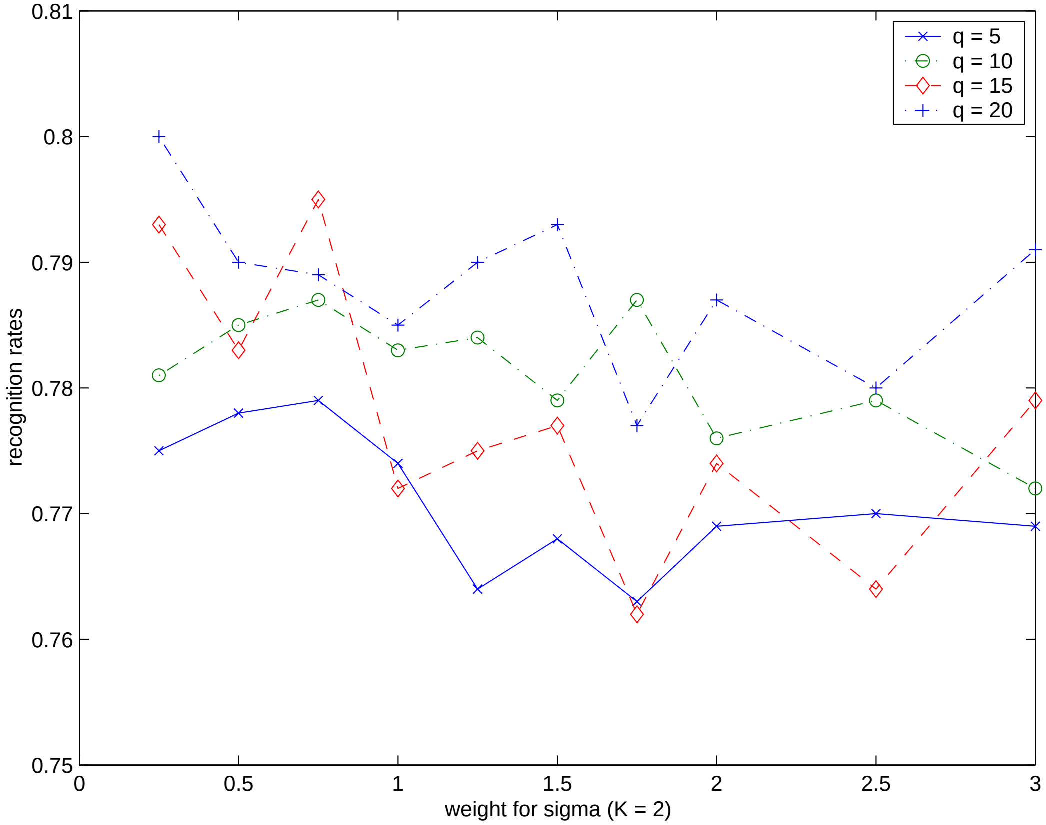


Figure 6.16: Recognition rates of the system using PWPCA clustering on the Cat and Dog data set for different weights applied to the σ. For $K = 2$, 5, 10, 15, and 20 subspace dimensions are tested.

designed as a combination of classification and exemplar matching components. We compared the system's overall performance with the performance of its components alone.

For the 2D synthetic data set generated by randomly sampling points from two Gaussian distributions, the system did not perform as well as the classification component with clustering algorithms that are able to fit the underlying distributions almost perfectly. However, the system's performance is not much lower and it still improves the performance over exemplar matching component – global PCA – alone.

To test the system on real images, we formed three data sets containing images of cats and dogs, and different types of objects extracted from aerial images. On these data sets, the experiments were conducted for average complex edge magnitude and Hough space features with different values for number of clusters and subspace dimensions. We also showed the influence of balancing in-subspace distance and outer-subspace distance in probability computation for the PWPCA version of the system.

The results showed that designing a recognition system using a combination of classification and exemplar matching components can perform more effectively than any of its components alone. Although the relative performance of the system and the global PCA varies depending on K and q, regardless of the features used, the system as a whole outperforms or is at least comparable to its component subsystems. More particular, the K-Means version of the system performs very effectively most of the time for the FH1 and FH2 data sets. The fact that our system works particularly well for small q's indicates that the system is more beneficial when the stored data is highly compressed.

Chapter 7

Supplementary Studies on Subspace Projection Algorithms

As described in Section 5.3, exemplar matching is implemented by subspace projection algorithms. Currently, only PCA was used in the proposed system because PCA is considered a standard technique for subspace projection and is more stable and robust than the other algorithms described in Section 5.3. Psychological and anatomical data, however, give no reason to prefer PCA over other subspace projection techniques. We performed comparative studies on subspace projection algorithms in isolation of the rest of the proposed system to provide an insight for their effectiveness in practice. The work presented in this chapter has its own importance in the machine vision community, where researchers have recently studied this subject, and provides some future directions for expanding the current system.

In the proposed system, a unique subspace is extracted for every cluster formed by the categorization subsystem. To compare subspace projection algorithms, we assume that the input data set is a single class cluster, i.e. a set of data samples already grouped by the clustering process so that they look alike. Hence, we performed the evaluation with a set of face images. Faces not only compose a single class data set, but are a well-known domain that humans become expert with. Moreover, many of the previous comparative studies on subspace projection algorithms were performed for face recognition, so our work can be directly compared with those previous results. Indeed, experimental results described in this section verify the results of previous comparisons in the literature.

7.1 The FERET Face Database

The FERET face recognition database is a set of face images collected by NIST from 1993 to 1997[1]. In the following studies, we used head-on images only, with 1,196 gallery images, 501 training images and four different sets of probe images. Using the terminology in [107], the *fb* probe set contains 1,195 images of subjects taken at the same time as the gallery images. The only difference is that the subjects were told to assume a different facial expression than in the gallery images [2]. The *duplicate I* probe set contains 722 images of subjects taken between one minute and 1,031 days after the gallery image was taken. The *duplicate II* probe set is a subset of 234 duplicate I probes, where the probe image is taken at least 18 months after the gallery image. Finally, the *fc* probe set contains 194 images of subjects under significantly different lighting conditions.

The face images in the database are pre-processed through a normalization process which produces images in the format used by NIST in FERET studies. In the process, the images are first registered by the eye positions and cropped to a smaller size so that only the face is included. Then, the backgrounds, such as hair and clothes, are removed by applying an oval-shaped mask to each image and histogram equalization is performed on the resulting non-masked pixel values. Finally, each image

---

[1] http://www.itl.nist.gov/iad/humanid/feret/

[2] Subject were not told what type of facial expression to assume for either the gallery or fb images, only that the two expressions should be different.

is standardized over non-masked pixels so as to have zero mean and unit standard deviation. Examples from the FERET database are shown in Figure 7.1.

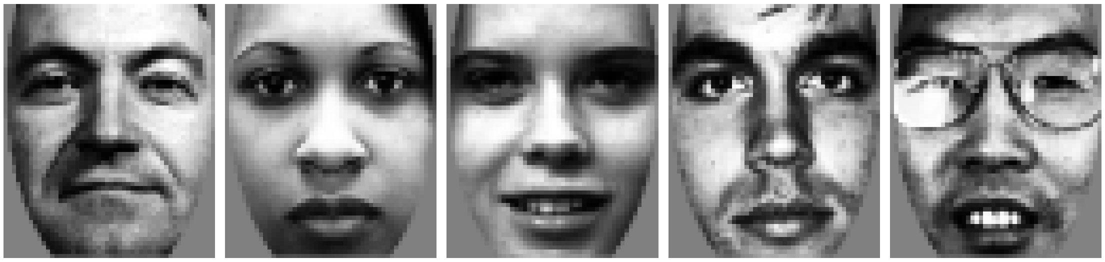


Figure 7.1: Sample images from the FERET database.

7.2 PCA vs. FA

In this section, we compare PCA and FA on a standard face recognition task. The results show that PCA significantly outperforms FA. In the second experiment, however, we show that the unique variances estimated by FA can be used to automatically detect and suppress background pixels prior to the application of PCA, and thereby improve the performance of PCA-based object recognition systems. This work is also presented in [5].

7.2.1 Recognizing Facial Identities

While PCA has been a popular method for face and object recognition, FA has not been widely exploited for object recognition. Here, a comparison between PCA and FA is made on a standard face recognition task using images from FERET database. To run FA, the 150×130 images were scaled down to 24×21 pixels[3]. The recognition process is as follows. First, a set of subspace basis vectors are computed for both

[3]When the dimension of input data is p, FA requires the inversion of $p \times p$ matrix. To avoid inverting singular matrices, we scaled down the image size. For comparison, we also used scaled images for PCA.

PCA and FA using 501 training images and the gallery images are projected into the subspace. For matching, a probe image is projected into the subspace and the closest gallery image to the probe image, as measured in the subspace, is retrieved. The FA model in equation 5.2 shows that there are $p(q+1)$ parameters when p dimensional data is represented by q factors. On the other hand, PCA only needs pq parameters in such case. Therefore, to make the number of parameters in FA and PCA the same, we used 200 factors and the first 201 (40%) principal components. The factor scores were computed as expected values conditioned on the observation [129]:

$$E[\mathbf{z}|\mathbf{x}_i] = (\mathbf{I} + \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} (\mathbf{x}_i - \mu)$$

Table 7.1 shows the recognition results for each probe set. Clearly, PCA always significantly outperforms FA, which explains why FA has not been used as a single global model for object recognition.

	PCA	FA
fb	1015 (84.93 %)	725 (60.67 %)
dup I	281 (38.92 %)	157 (21.75 %)
dup II	36 (15.38 %)	15 (6.40 %)
fc	63 (32.47 %)	9 (4.64 %)

Table 7.1: Performance of PCA and FA on different probe sets [5].

7.2.2 FA for Background Suppression

FA has been used to fit mixtures of local linear models [47, 63]. These systems exploit an Expectation-Maximization (EM) [35] algorithm for fitting mixtures of factor analyzers to data sets [47, 55]. On the other hand, PCA has been considered simply a transformation of the data. Recent work by Tipping and Bishop [148], however, derives a probabilistic model for PCA and extends it to a mixture of local PCA models. As a result, there are now methods to compute mixtures of PCA models as well as mixtures of FA models.

Is there any reason, then, to prefer FA over PCA for object recognition? We can find no comparisons in the literature, but the data presented earlier reflects our experience that PCA outperforms FA on standard face recognition tasks. There are circumstances, however, under which a combination of FA and PCA outperforms either alone. The key observation lies in the analysis by Tipping and Bishop, who show that the difference between FA and PCA lies in the residual error model [148]. PCA assumes that all variance not accounted for by the eigenvectors is drawn from a single zero-mean Gaussian distribution with standard deviation σ, and determines the principal components which can account for the total (both common and unique) variance. Therefore, PCA is highly sensitive to variation in pixel noise and the presence of background.

FA, on the other hand, fits a unique standard deviation σ_i to every pixel. As a result, the variance of a variable is split into common and unique variance in the FA model (equation 5.3 in Section 5.3.3). Typically, FA applications are concerned with the common variance and exclude the unique variance; however, we are more interested in the latter. The unique variances (i.e. the diagonal elements of $\mathbf{\Psi}$) determine how much of the variability in each input dimension is not attributable to linear factors. Therefore, when the input data is a set of 2D images, it models the independent pixel noise in the data set. We assume that non-linear variances are attributable to backgrounds since this model is applied to a set of independently collected images rather than a video sequence in which same backgrounds appear repeatedly. This implies that FA may have an advantage over PCA in circumstances where the foreground and background pixels are not separated *a-priori*, since background pixels should receive higher unique variances.

In fact, this is what happens for object recognition systems taking a set of images as input. Even though the images contain similar objects, the background may still be quite different, unless the images are preprocessed to separate background from foreground. It would severely affect the system's performance especially when PCA-like matching methods are used. The proposed system also has the same problem. A scene generally contains a variety of objects with different shapes and, as shown in the previous chapter, the input patch images are generated independently and no preprocessing is applied for background removal. Once they are grouped by clustering methods, each cluster includes images of similar objects, but they may have quite different backgrounds, which will consequently drop the performance of exemplar matching. Therefore, it is important to suppress background pixels before a subspace is fit to a cluster.

To test our hypothesis, we apply FA to two different versions of the FERET face database with different amounts of background. They are formed by cropping the images in the database to two different sizes: a standard size of 150×130 pixels (as in [107]), and 200×170 pixels. The left column of Figure 7.2 shows the images of one person from each data set. Obviously, the larger image includes more background, particularly hair and clothes. To run FA, the 150×130 images were scaled down to 24×21 pixels, while the 200×170 images were scaled to 25×21 pixels. The matrix $\mathbf{\Psi}$ is computed by FA on 1,301 training and gallery images. The diagonal elements form the variance images shown in the right column of Figure 7.2. As we can see, pixels outside the face and around the extreme points of some facial structures – eyebrows, nose and mouth – vary a lot across the data set, while areas of facial skin have lower unique variance. This is particularly true in the bottom right image, where the background is captured quite well by the unique variances.

In order to suppress background pixels relative to foreground pixels, source pixels in the images are weighted by the inverse of their unique standard deviation as estimated by FA. The recognition algorithm applied after weighting is PCA followed by a nearest neighbor classifier, as in [107]. To demonstrate the use of FA in suppressing background pixels, we compare the recognition performance of PCA on the weighted

Figure 7.2: The left column shows an example from the FERET database cropped into two different sizes. On the right, the variance map of the data sets of smaller sized images (top) and larger sized images (bottom) computed by applying FA to combined set of training and gallery images from each data set.

images (WPCA) to that of PCA performed on an unweighted data set.

Table 7.2 and 7.3 show the resulting recognition performance. The weighting process does not help for the smaller images, since they do not include much background to suppress. Most of the unique variance captures internal variations around the facial structures. Statistically[4], there is no significant difference in performance between the two methods, except on the fb probe set where PCA outperforms WPCA. This makes sense, since the fb images were taken immediately after the gallery images. The backgrounds (e.g. hair, clothes) are therefore the same for each individual in the gallery and fb data sets, so the background represents useful information rather than noise. In all instances where the background is different between the probe and

[4]To test statistical significance, we apply McNemar's pairwise-difference test.

gallery, WPCA performs as well as PCA.

	PCA	WPCA
fb	1015 (84.93 %)	989 (82.76 %)
dup I	281 (38.92 %)	278 (38.50 %)
dup II	36 (15.38 %)	36 (15.38 %)
fc	63 (32.47 %)	67 (34.54 %)

Table 7.2: Performance of PCA and WPCA on different probe sets. The original image size of the data set is 150 x 130 pixels [5].

However, with larger images that include more background, WPCA outperforms PCA except for the fc probe set. For the fb, dup I, and dup II probe sets, WPCA performs significantly better than PCA with over 99% confidence. Compared to the results on the set of smaller images shown in Table 7.2, this shows that the proposed weighting process, using the unique variances captured by FA, helps suppress the background. As for the fc probe set, it is not clear what causes the anomalous result. The fc probe set is the smallest probe set, and the lighting for the fc images is from a different direction than in the other three probe sets. The fc images are also darker than the other images, so there is less dynamic range in the source pixels. Why any of these factors should differentially effect WPCA and PCA remains for further investigation.

	PCA	WPCA
fb	1046 (87.53 %)	1087 (90.96 %)
dup I	285 (39.47 %)	308 (42.66 %)
dup II	30 (12.82 %)	40 (17.09 %)
fc	108 (55.67 %)	93 (47.94 %)

Table 7.3: Performance of PCA and WPCA on different probe sets. The original image size of the data set is 200 x 170 pixels [5].

It can be argued that using FA for background suppression seems to have an inherent contradiction when it is applied to face recognition as described in this

Section. FA captures high variance in pixels around facial structures (Figure 7.2), which are often considered prominent features in face recognition, and the proposed process seems to suppress these features. The results in Table 7.2, however, show that suppressing these pixels does not lower performance; there is no significant difference between PCA and WPCA. It may be that movable features are less predictable than fixed features that are not suppressed, or it may be because face recognition is more sensitive to holistic or configural properties than localized features [23]. Indeed, in the next section, we show that localized features may be more important for analyzing facial expressions [8] than recognizing individual identities.

7.3 PCA vs. ICA

Compared to PCA, it is quite recently that ICA has been applied to image analysis [16], recognizing faces [9, 90, 105, 159] and expressions [39]. Since then, a number of comparisons between PCA and ICA have been made. The comparisons were performed mostly in the context of face recognition system and the results are somewhat contradictory. Bartlett, et al. [9], Liu and Wechsler [90], and Yuen and Lai [159] claim that ICA outperforms PCA for face recognition, while Baek et al. [6] claim that PCA outperforms ICA and Moghaddam [105] claims that there is no statistical difference in performance between the two. The relative performance of the two techniques is therefore an open question. Adding to the confusion are the two different architectures for applying ICA to recognition tasks as described in Section 5.3.2. In this section, the space of PCA and ICA comparisons is explored by systematically testing two ICA architectures and three PCA distance measures. This work is previously reported in [6] and [41].

7.3.1 Recognizing Facial Identities

The baseline face recognition system used in this comparison is the same as describe in Section 7.2. This time, we scaled down the face images to 60×50 pixels to run ICA and PCA is run on the same sized images. The training images are a randomly selected subset of 500 gallery images.

In this experiment, we kept 200 subspace dimensions (40% of the maximum possible number of non-zero eigenvalues given a training set of 500 images). Rather than pick a single distance measure, we tested PCA three times, once using the L1 (city-block) distance metric, once using the L2 (Euclidean distance) metric, and once using a Mahalanobis (L2 after scaling each dimension by the square root of its eigenvalue) distance metric. ICA was tested also keeping 200 basis vectors and a cosine distance measure was selected to retrieve images in the ICA subspaces[5]. The other InfoMax parameters for architecture I were a block size of 50, and a learning rate that began at 0.001 and was annealed over 1,600 iterations to 0.0001. The parameters for architecture II were the same, except that the learning rate began at 0.0008 rather than 0.001 because of numerical limitations. These parameters match those used in [9].

Table 7.4 shows the recognition rate for PCA, ICA architecture I, and ICA architecture II, broken down according to probe set. Results for PCA are divided according to whether the L1, L2 or Mahalanobis distance metric was used. The most striking feature of Table 7.4 is that ICA architecture II always has the highest recognition rate. PCA is second, with both the L1 and Mahalanobis distance measures performing well. The performance of ICA architecture I and PCA(L2) are very close, and neither is competitive with architecture II or PCA with L1 or Mahalanobis. According to Mc-

[5]We also tested L1, L2, and Mahalanobis distance measures. For ICA architecture II, the cosine measure clearly outperforms these other measures. For architecture I, cosine clearly outperforms L1 and L2 and there is no significant difference between cosine and Mahalanobis.

Nemar's significance test, the difference in performance between architecture II and PCA(L1) is significant for all four probe sets at a confidence level of 97% or higher.

probe set	ICA (cosine)		PCA		
	Arch. I	Arch. II	L1	L2	Mahalanobis
fb	73.72 %	82.26 %	80.42 %	72.80 %	73.23 %
dup I	36.15 %	48.48 %	40.30 %	33.24 %	39.34 %
dup II	14.53 %	32.48 %	22.22 %	14.53 %	24.36 %
fc	5.67 %	51.03 %	20.62 %	4.64 %	39.69 %
average	50.62 %	64.31 %	57.31 %	49.17 %	56.16 %

Table 7.4: Recognition rates for PCA and both architectures of ICA on the FERET face database. The task is to match the identity of the probe image [41].

The results in Table 7.4 were generated using 200 subspace dimensions. In general, the relative ordering of the subspace projection techniques does not depend on the number of subspace dimensions. Figure 7.3 and Figure 7.4 plot the recognition rate for all five techniques as a function of the number of basis vectors. For most techniques, the lines never cross. The one exception is PCA with the Mahalonobis distance metric, which performs almost as well as ICA architecture II with small numbers of subspace dimensions, but whose performance drops off relative to ICA architecture II and sometimes relative to PCA with L1 as the number of subspace dimensions increases.

In many cases, the results shown in Table 7.4 do not contradict the previous results in the literature, if the ICA architecture and distance measures are taken into account. For example, Baek et al. [6] found that PCA with the L1 distance measure outperformed ICA architecture I, which is consistent with Table 7.4. Liu and Wechsler [90] compared ICA architecture II to PCA with L2, and found that ICA was better. Again this agrees with Table 7.4, even though on the surface it appears to contradict the results in Baek, et al. Similarly, Bartlett et al. reported in 1998 that ICA architecture II outperformed PCA with L2 [9, 10], a result predicted

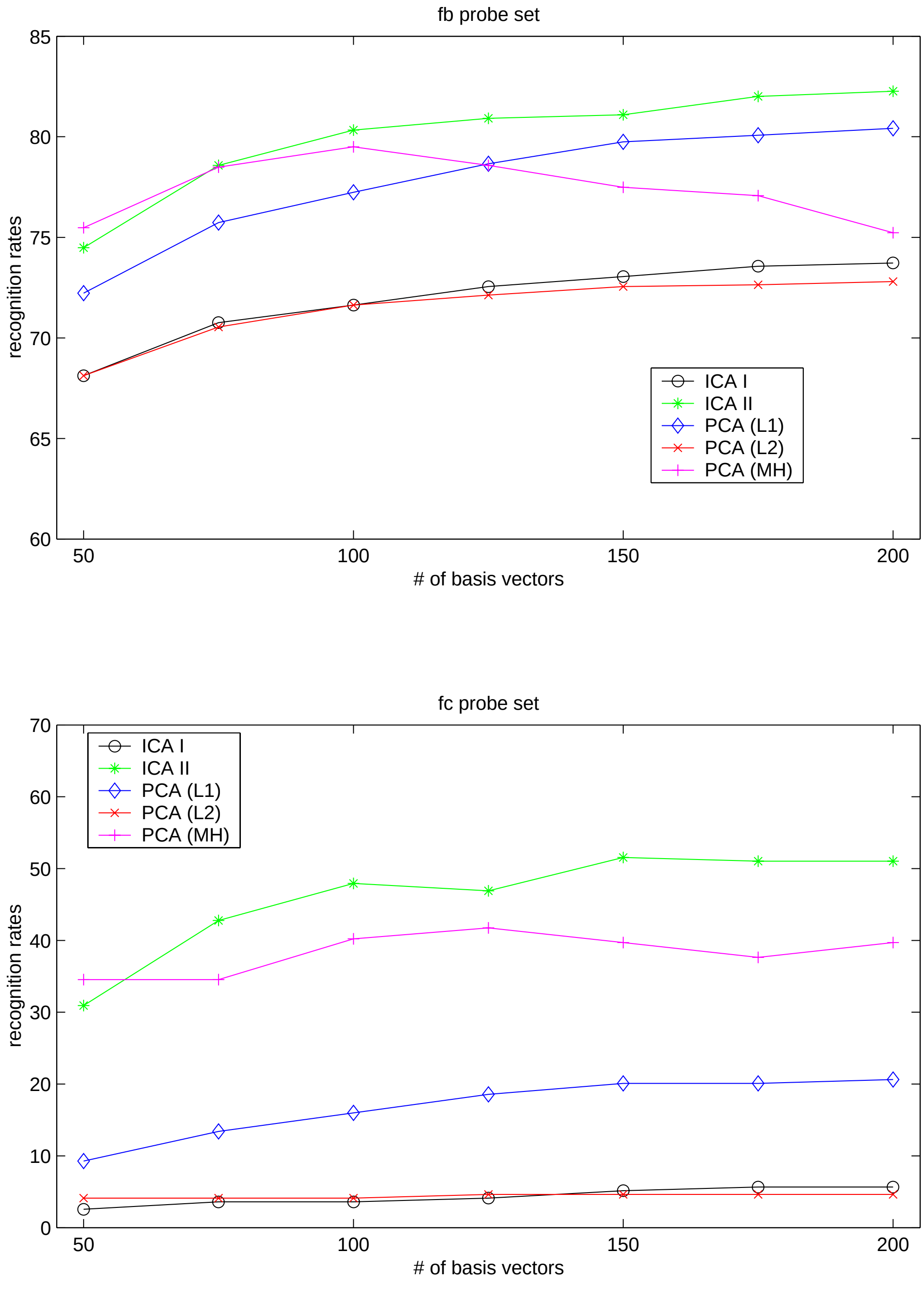


Figure 7.3: Recognition rates for ICA Architecture I (black), ICA Architecture II (green), and PCA with the L1 (blue), L2 (red) and Mahalanobis (magenta) distance measures as a function of the number of subspace dimensions. Top graph corresponds to fb probe set and the bottom graph corresponds to fc probe set. Recognition rates were measured for subspace dimensionalities starting at 50 and increasing by 25 dimension up to a total of 200 [41].

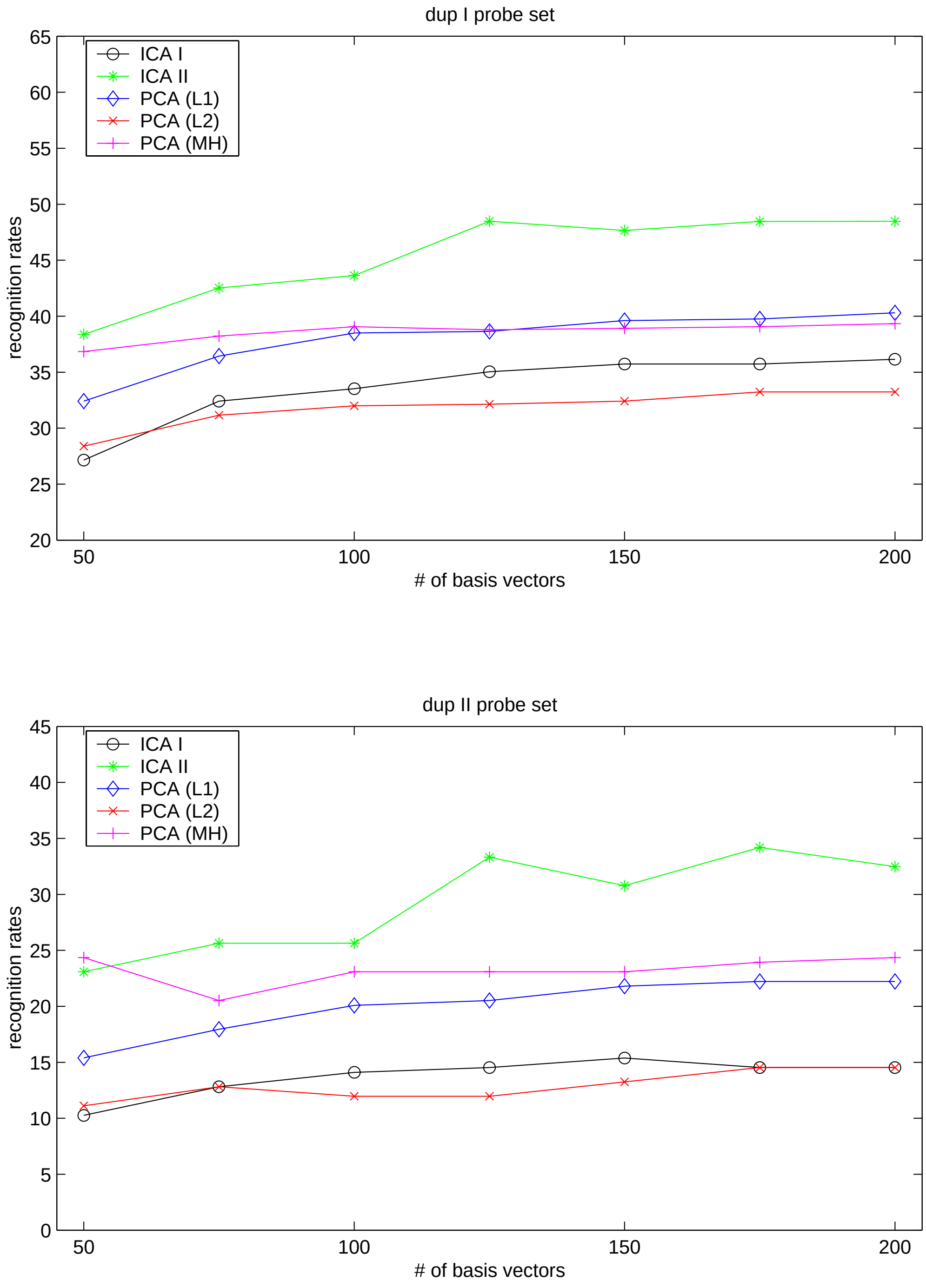


Figure 7.4: Recognition rates for ICA Architecture I (black), ICA Architecture II (green), and PCA with the L1 (blue), L2 (red) and Mahalanobis (magenta) distance measures as a function of the number of subspace dimensions. Top graph corresponds to dup I probe set and the bottom graph corresponds to dup II probe set. Recognition rates were measured for subspace dimensionalities starting at 50 and increasing by 25 dimension up to a total of 200 [41].

by Table 7.4.

The breakdown of results in Table 7.4, however, reconciles results previously reported in the literature that appeared contradictory. It shows that appropriate distance measure and ICA architecture have to be used for comparison. Two of the papers mentioned above compared ICA architecture I to PCA with the L2 distance metric, and found that ICA outperformed PCA [9, 159]. Our study also found that ICA architecture I was slightly better than PCA with the L2 distance metric, but our results were only statistically significant for the dup I probe set. With different training and test images, these studies may have detected a significant difference where we did not. Our data does not contradict their results, although it fails to find strong support for them. We think, however, that the previous conclusions should be limited to the architectures and distance metrics that they tested.

The one previous result that contradicts ours is the study by Moghaddam [105] which found no statistically significant difference between ICA architecture II and PCA with the L2 distance metric. According to Table 7.4, ICA architecture II should have won this comparison, and the result should not have been in doubt. From the paper, however, it appears that Moghaddam may have used the L2 distance metric for ICA as well as for PCA. This significantly lowers the performance of ICA architecture II[6] and may explain part of the discrepancy.

7.3.2 Recognizing Facial Actions

In comparisons between PCA and ICA on face recognition tasks, we showed that the superiority of one algorithm over another is not absolutely decidable, but rather depends on the distance measures for matching and the ICA architecture types to be

[6]The average recognition rate of ICA architecture II with the L2 distance metric dropped to 55.32%, a 9% drop in accuracy. But ICA still performs better than the PCA with L2 distance.

applied. In this section, we expand the study to a different domain of task: facial action recognition. We show that the relative performance of the two techniques also depends on the nature of given task.

7.3.2.1 The Facial Action Database

The data set, called the *Ekman-Hager data set of directed facial actions*, consists of images of subjects performing specific facial actions. In particular, Ekman has described a set of 46 human facial actions [44], and Paul Ekman and Joe Hager have collected images of people performing these actions. They asked 20 subjects to perform six facial actions each. Not all subjects were capable of performing all actions in isolation of others, so only 80 total subject/action pairs were collected[7]. A temporal sequence of five images was taken of each subject performing each action, where the first image shows only small movements at the beginning of action. The movements become successively more pronounced in images 2 through 5. The data set also contains mirror-reversed examples of each action in order to boost the amount of training data. All five images are then subtracted from the picture of the same subject with a neutral expression, prior to the start of the facial action. The resulting data is composed of difference images, as shown in Figure 7.5. This data is further described in [8]. For the experiments in this study, only the upper halves of the faces were used.

7.3.2.2 Recognition Results

The appropriate distance measure and ICA architectures are also dependent on the nature of the given task. In this Section, we compare the performance of PCA and ICA on recognizing facial actions rather than facial identity. The motivation is that

[7]For the Action Units 1, 2, 4, 5, 6, and 7, there are 9, 10, 18, 20, 5, and 18 subjects, respectively, who were able to perform each action, which make total of 80 subject/action pairs.

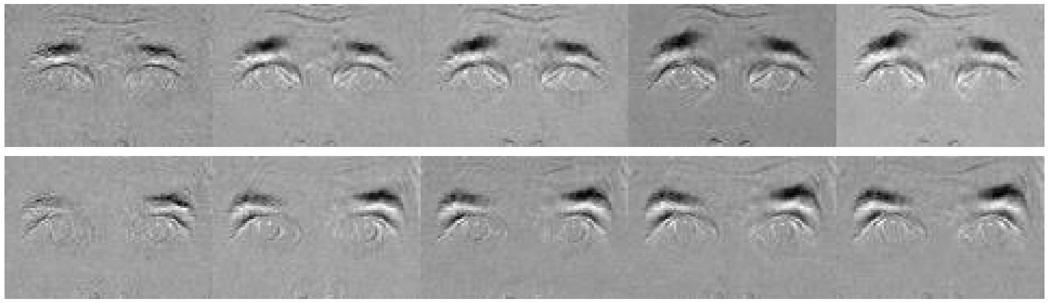


Figure 7.5: Sequences of difference images for Action Unit 1 and Action Unit 2. The frames are arranged temporally left to right, with the left most frame being the initial stage of the action, and the right most frame being its most extreme form [41].

recognizing expressions is a significantly different task from recognizing identity, as suggested in [8, 39]. Identity recognition is hypothesized to be a configural (as opposed to spatially localized) task [23], therefore it should be modeled better by spatially overlapping (global) basis vectors. Our data presented in Section 7.3.1 suggests that ICA architecture I, which produces spatially localized basis vectors (middle row in the Figure 5.8), performs worse than either PCA or ICA architecture II, both of which produce global basis vectors (top and bottom rows in the Figure 5.8), and thereby supports the configural view of facial identity recognition. On the other hand, facial actions are formed by moving localized muscle groups, therefore it seems reasonable that localized features would form a better basis for recognizing facial actions than spatially overlapping features. Here we show how the recognition performance is influenced by the localized properties of the given task.

For testing purposes, the subjects in the Ekman-Hager data set were partitioned into four sets. The four sets were designed to keep the total number of actions as even as possible, given that not all subjects performed all actions. This partition by subject index is given in the Table 7.5. Otherwise, the methodology and parameters were the same as described in Section 7.3.1, except that in this experiment a trial was a success if the retrieved gallery image was performing the same facial action as the probe image. We performed the experiments twice, once restricting the probe set to only the third image in each temporal action sequence (as in [8]), and once allowing all five images to be used as probes. Once again, ICA architecture I, ICA architecture II, and PCA with the L1, L2 and Mahalanobis distance metrics were compared. Because of numerical limitations in our ICA architecture II implementation, only 110 features were computed for each technique.

action	partition (subject #)			
unit #	1 (0,2,4,5,14)	2 (1,7,8,9,16)	3 (3,10,11,12,13)	4 (6,17,18,19,20)
1	2	2	2	3
2	2	3	2	3
4	5	4	4	5
5	5	5	5	5
6	1	1	2	1
7	4	4	5	5
total	19 actions	19 actions	20 actions	22 actions

Table 7.5: Subject partition. Each row corresponds to a facial action, each column to a set of subjects. Table entries correspond to the number of subject/action pairs in a partition for the corresponding facial action [41].

One difference between ICA and PCA is that the basis vectors in ICA are not ranked in order. To compare the performance for subsets of basis vectors, class discriminability r is used to order the basis vectors in ICA[8] [8]. To compute r, each training image is assigned a facial action class label. Then, for each feature, the between-class variability $\sigma_{between}$, and within-class variability, σ_{within} of the corresponding coefficients are computed by:

$$\sigma_{between} = \sum_i (\mu_i - \mu)^2$$

[8]ICA basis vectors were not ordered by relevance for the identity recognition task because the FERET data set contains only two head-on images of most subjects, and no more than four head-on images of any subject.

$$\sigma_{within} \quad = \quad \sum_i \sum_j (b_{ij} - \mu_i)^2$$

where μ is the overall mean of a coefficient across the training images, μ_i is the mean for class i, b_{ij} is coefficient of j^{th} training image in class i, and r is the ratio of $\sigma_{between}$ to σ_{within}, i.e. $r = \sigma_{between}/\sigma_{within}$. This allows us to rank the ICA features in terms of discriminability and to plot recognition rate as a function of the number of subspace dimensions. We create the same plot for PCA, ordering its features according to the eigenvalues (Figure 7.6)[9].

The results of this experiment are a little more complex than in Section 7.3.1, in part because recognition rate did not increase monotonically with the number of subspace dimensions. Figure 7.6 shows the recognition rate as a function of the number of dimensions for each of the five techniques (ICA architecture I, ICA architecture II, and PCA with L1, L2 and Mahalanobis), averaged across all four test sets. For all five techniques, the maximum recognition rate occurs with a different number of subspace dimensions on every test set, suggesting that it may not be possible to tune any of the techniques by changing the number of subspace dimensions. Table 7.6 therefore presents the recognition rates for every technique using 110 basis vectors. For the same results when only the third images from each temporal sequence are used as probes, see [41]. (The results are essentially equivalent.)

When recognizing facial actions, it is ICA architecture I that outperforms the other four techniques at a confidence level of 99% or higher. PCA(L2) is second best, followed by PCA(L1), ICA architecture II, and PCA(Mahalanobis). This is consistent with the hypothesis that spatially localized basis vectors outperform spatially overlapping basis vectors for recognizing facial actions, and underscores the point that the analysis technique must be selected based on the recognition task. It is also

[9]We also tested PCA using relevance ordering, but its performance was better using the ordering of the eigenvalues.

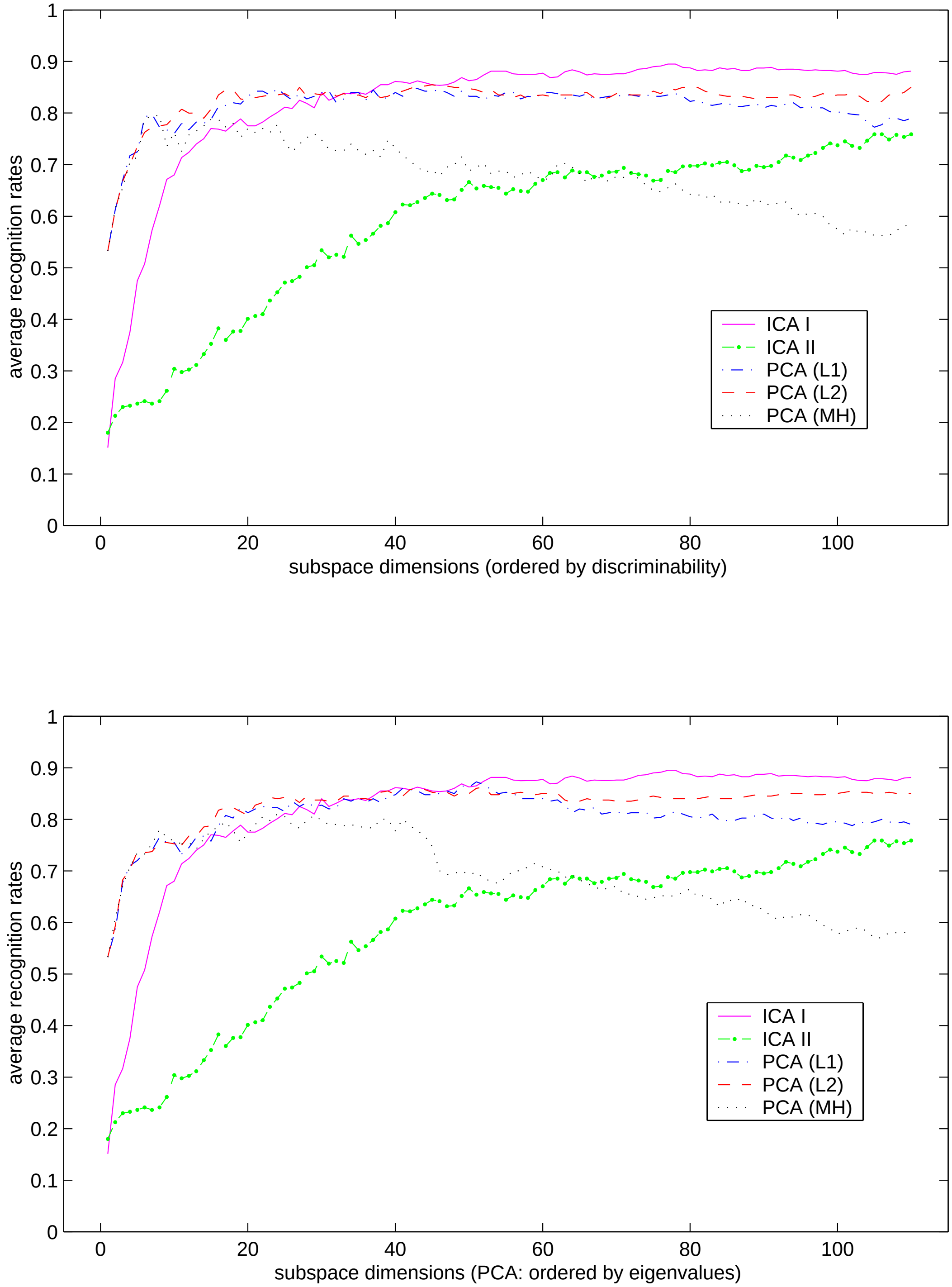


Figure 7.6: Recognition rates vs. subspace dimensions. On the top, both ICA and PCA components are ordered by the class discriminability while PCA components are ordered according to the eigenvalues in the bottom plot. ICA architecture I is magenta, ICA architecture II is green, PCA with L1 is blue, PCA with L2 is red, PCA with Mahalanobis is black [41].

test set	ICA (cosine)		PCA		
	Arch. I	Arch. II	L1	L2	Mahalanobis
1	87.37 %	74.73 %	83.16 %	83.16 %	52.63 %
2	95.79 %	84.21 %	84.21 %	92.63 %	70.53 %
3	82.50 %	69.00 %	80.00 %	83.00 %	53.00 %
4	87.27 %	75.91 %	70.00 %	81.82 %	58.18 %
average	88.12 %	75.87 %	79.00 %	85.00 %	58.50 %

Table 7.6: Recognition rates for facial actions using PCA and both architectures of ICA. The images were divided into four sets according to Table 7.5 and evaluated using 4-fold cross validation. Techniques were evaluated by testing from 1 to 11-subspace dimensions and taking the average [41].

consistent with the data in [8, 39]

7.3.3 Discussions

Comparisons between PCA and ICA are not simple, because no single technique is always best. Differences in parameters, architectures, distance metrics, and given tasks must be taken into account. Experiments presented in this section explore the space of PCA/ICA comparisons, by systematically testing two ICA architectures and three PCA distance measures. In the process, we were able to verify the results of previous comparisons in the literature and to relate them to each other. Of particular interests, the nature of given tasks influence the performance of different ICA architectures. ICA architecture I is suitable for tasks with localized properties while ICA architecture II is better for configural tasks.

As with any comparative study, there are also limitations. The space of InfoMax parameters were not explored because of its size and run-time constraints. Instead, parameters were selected based on previous experience with the algorithm [9, 8, 7, 39]. It is always possible that another choice of parameters might have improved the performance of ICA. By testing only InfoMax, we did not explore the space of ICA algorithms, either. However, others have reported little difference between different

ICA algorithms [31, 67].

A more serious limitation is that PCA and ICA were evaluated according to only one criterion, recognition rate. Other criteria, such as computational cost, may apply in some circumstances. PCA has also been demonstrated to be effective across a wide range of task domains while ICA is less thoroughly investigated.

7.4 Summary

In this chapter, we presented supplementary studies on subspace projection algorithms for face recognition tasks. Face images taken from a same viewpoint compose a single class data set, which can be assumed a cluster output from the categorization subsystem. Since a subspace is extracted for each cluster independently, the evaluation done here should apply to the proposed system. Faces are also a well-known domain that humans become an expert with, and the literature provides many comparative studies on subspace projection algorithms for face recognition tasks. Our results were compared with those previously reported evaluations.

It is shown that FA does not do well for recognizing objects by itself. The linear factors computed by FA did not outperform the PCA basis vectors, however, the resulting unique variance image separated background from foreground pixels quite well. We used the unique variance to inversely weight pixels prior to applying PCA. This suppressed background pixels relative to foreground pixels, and improved the performance of PCA. Therefore, FA can be applied to automatically suppressing backgrounds prior to the recognition process. The method presented here has the property that it improves recognition performance when background is present in the images, but does not significantly reduce performance when no or little background is present.

Comparisons between PCA and ICA were done by exploring ICA architectures, distance measures for matching, and different nature of given tasks. The results show that no single technique is always the best, and the relative performance of the two techniques is depending on many factors. Especially, the performance of different ICA architectures is influenced severely by the nature of given tasks, while PCA maintains relatively stable performance.

The experimental results presented in this chapter are not yet embedded in the proposed system. However, the tasks evaluated here are important in the context of the current system. At this moment, PCA is the only technique used for exemplar recognition and there is no explicit background removal mechanism implemented in the system. Including different architectures of ICA as system options and applying FA prior to exemplar matching to suppress backgrounds would improve the system's performance.

Chapter 8

Conclusions

Understanding the mechanisms underlying visual object recognition has been an important subject since the early days of cognitive science. Researchers in various areas have made tremendous efforts to provide a general model for biological vision systems and to theorize computational models for building artificial vision systems. Although still far from finding a complete model for biological vision systems, results from many studies indicate that a central principle that characterizes vision is functional specialization. It has been generally considered that the ventral visual pathway plays the critical role in identifying and recognizing objects. Furthermore, it has been proposed that there are separate visual areas dedicated to recognizing objects with which humans become expert.

In this work, we have tried to provide a possible algorithmic analysis of psychological and anatomical models of the ventral visual pathway, more specifically the pathway that is responsible for expert object recognition, using the current state of machine vision technology. As a result, we propose an expert object recognition system based on two biological theories: Kosslyn's model of higher-level visual processing and Tarr and his colleagues' work on viewpoint-dependent mechanism and perceptual expertise for visual object recognition. Kosslyn's model provides a design framework for the overall structure of the system, while Tarr's work provides a theoretical context for training and testing the system.

The proposed system is composed of multiple, functionally distinct components performing feature extraction and pattern matching. When evaluating the system, we focus on the last two components that are responsible for matching input patterns to those stored in memory. They correspond to classification and exemplar matching and each of the components alone can be considered as a standard object recognition system. The main purpose of the experiments conducted in this work was to analyze the system's effectiveness as designed by a combination of the two components.

The proposed system is tested with four different multi-class data sets. The overall performance of the system is compared to the performance of classification and exemplar match alone. In general, the system's performance varies depending on the features used in the pattern matching component and the parameter values that affect each component's behavior. However, for all data sets, the results showed that the system as a whole outperforms any of its component subsystems when the number of sub-dimensions used for matching is relatively small. This indicates that the system can be more effective when the data is highly compressed. Performance of the system also changes according to the clustering algorithms. Currently, K-Means produces better performance than the PWPCA clustering, which has many more parameters than K-Means that have to be tuned.

Even though the system was not tuned to behave optimally for different input data sets, it performs better than or comparable to the component subsystems. It should be noted that the exemplar matching subsystem alone implements global PCA, which is one of the most-studied and the best performing techniques for 2D object recognition. The experimental results show that a recognition system designed based on biological models described in this work matches the performance of a single state-of-the-art machine vision technique and can perform even better when it is appropriately tuned.

8.1 Contributions

The main contribution of this work is a system that implements a biological model of expert object recognition. The system provides explicit connections between computational and biological models of visual object recognition at the level of major components. It is also a complete end-to-end system, in which every component is based on a biological model. Each of the computational approaches implementing the different levels of recognition for pattern matching in the system can be considered as a standard object recognition system by itself. Combining them into one framework is shown to be more effective than using the components alone, and it also enables the system to perform recognition in different levels, which is a defining characteristic of expert object recognition.

The system also provides a baseline mechanism for testing Kosslyn's biological model of object recognition. The system can be used to test Kosslyn's model with different computational alternatives and to be extended to a more complex, biologically plausible vision system. Some connections in Kosslyn's model, however, were not implemented in the proposed system. For example, in Kosslyn's model, there is a connection from the exemplar pattern activation subsystem to the category pattern activation subsystem and a direct connection from the preprocessing subsystem to the exemplar pattern activation subsystem. While the latter is implemented indirectly since it is equivalent to setting the number of clusters to one, the connection from the exemplar subsystem toward the category subsystem did not exist in the current system. It is hard to fit this connection with the algorithms used for implementing the two subsystems. This demands a clarification of the role of the connection from the psychological community so that more suitable algorithms can be considered. Feedback like this can facilitate interactions between computational and theoretical fields of study in vision.

While implementing the system, we developed a clustering algorithm, which approximates traditional EM for fitting a Gaussian Mixture model to the data set. Our algorithm reduces the dimensionality limitation of the traditional EM to some extent. We also conducted a thorough comparative evaluation on the subspace projection algorithms. It provides important feedback to the machine vision community, where many researchers have studied the subject. Our work resolved somewhat contradictory results previously reported in the literature. In addition to the performance evaluation, we also proposed an alternative way of using FA for background suppression. Although these studies on subspace projection algorithms were conducted independently from the context of the proposed system, they provide valuable insights for future directions to expand the current system.

8.2 Future Work

Many issues for future work are suggested by the work presented in the previous chapters. Among them are the following:

- **Include ICA for exemplar matching:** The experiments described in Section 7.3 showed that ICA can outperform PCA if an appropriate architecture is applied. Therefore, including the two architectures of ICA into the system provides more options for exemplar matching and might improve the system's overall performance.

- **Include FA for background suppression:** As discussed in Section 7.2, the unique variances computed by FA separates background from foreground pixels quite well. Computing the unique variance image for each cluster and inversely weighting pixels by the variances prior to exemplar matching would have the effect of automatically suppressing backgrounds, thereby improving exemplar recognition performance.

- **Evaluate clustering algorithms:** Currently, the clustering algorithms used for classification are evaluated by the final exemplar matching performance. As shown in Section 6.1.2.3, clustering performance is not well measured by comparing samples to the dominant label in their cluster. As with the comparative studies performed for subspace projection algorithms, we need to develop a way to evaluate clustering algorithms in terms of grouping visually similar images.

- **Include imagery feedback model:** In Kosslyn's model of visual object recognition shown in Figure 3.1, there is a backward connection from the pattern recognition subsystem to the visual buffer, called imagery feedback. Kosslyn says that, when the recognition task is difficult, this top-down process maps the current hypothesized representation back to the visual buffer by continuously adjusting it until the feedback augments the input as well as possible. In practical systems, one way to model imagery feedback is to reconstruct strongly matched images in the memory and, rather than just filling in the input, compare them with the input for further analysis and performance improvement. In [147], imagery feedback is implemented as PCA followed by perceptron learning for difference estimation and it is shown that the recognition performance can be improved by the post-PCA processing.

- **Model self adaptation capabilities:** There are many factors that influence the system's behavior for given tasks. As discussed in Chapter 6, the system does not always perform effectively; its performance depends on the parameter values used. If we can make the system adaptive to given tasks, it will provide a far more effective artificial vision system that is biologically inspired. However, this work would require effort from both the machine and biological vision communities. A computational model should be provided based on the studies about the underlying mechanisms for the adaptive capabilities of biological vision systems, and we have to find a technical way to embed the model in the current system's framework.

Appendix A

EM Algorithm for Factor Analysis

In this appendix, we review the derivation of an EM algorithm for FA described in [55].

The generative factor analysis model is defined as:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{u}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$.

$$\begin{aligned}
\mathbf{p}(\mathbf{z}) &= (2\pi)^{-k/2} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right) \\
\mathbf{p}(\mathbf{x}|\mathbf{z}) &= (2\pi)^{-p/2}|\mathbf{\Psi}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{W}\mathbf{z})^T\mathbf{\Psi}^{-1}(\mathbf{x} - \mathbf{W}\mathbf{z})\right\}
\end{aligned}$$

Then,

$$\begin{aligned}
Cov(x) &= E(\mathbf{x}\mathbf{x}^T) \\
&= E\{(\mathbf{W}\mathbf{z} + \mathbf{u})(\mathbf{W}\mathbf{z} + \mathbf{u})^T\} \\
&= E(\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T) + E(\mathbf{W}\mathbf{z}\mathbf{u}^T) + E(\mathbf{u}\mathbf{z}^T\mathbf{W}) + E(\mathbf{u}\mathbf{u}^T) \\
&= \mathbf{W}E(\mathbf{z}\mathbf{z}^T)\mathbf{W}^T + \mathbf{W}E(\mathbf{z}\mathbf{u}^T) + E(\mathbf{u}\mathbf{z}^T)\mathbf{W} + E(\mathbf{u}\mathbf{u}^T) \\
&= \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}
\end{aligned}$$

Therefore, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \mathbf{\Psi})$.

$$\mathbf{p}(\mathbf{x}) = (2\pi)^{-p/2}|\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^T(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}\mathbf{x}\right\}$$

Using the joint normality of data and factors, the expected value of factors given $\mathbf{x}$ is computed by:

$$E(\mathbf{z}|\mathbf{x}) = \alpha\mathbf{x} \tag{A.1}$$

where $\alpha = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}$.

Also, the second moment of the factors given $\mathbf{x}$ is computed by:

$$\begin{aligned} E(\mathbf{z}\mathbf{z}^T|\mathbf{x}) &= \mathbf{C}_{\mathbf{z}|\mathbf{x}} - E(\mathbf{z}|\mathbf{x})E(\mathbf{z}|\mathbf{x})^T && \text{(A.2)} \\ &= \mathbf{I} - \alpha\mathbf{W} + \alpha\mathbf{x}\mathbf{x}^T\alpha^T && \text{(A.3)} \end{aligned}$$

Then, the log-likelihood of $\mathbf{x}$ is:

$$\begin{aligned} Q &= \log\left[\prod_{i=1}^{n}(2\pi)^{-p/2}|\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{x}^T(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}\mathbf{x}\right\}\right] \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}| - \frac{n}{2}\mathbf{x}^T(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}\mathbf{x} \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}| - \frac{n}{2}trace(\mathbf{C}_{\mathbf{x}}(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}) \end{aligned}$$

where $\mathbf{C}_{\mathbf{x}}$ is the sample covariance matrix defined by $\mathbf{x}\mathbf{x}^T$. $\mathbf{W}$ and $\mathbf{\Psi}$ are estimated so that the log-likelihood is maximized.

E-Step : Compute $E(\mathbf{z}|\mathbf{x}_i)$ and $E(\mathbf{z}\mathbf{z}^T|\mathbf{x}_i)$ for each $\mathbf{x}_i$, given $\mathbf{W}$ and $\mathbf{\Psi}$ using equations A.1 and A.3. Since $\mathbf{\Psi}$ is diagonal, the following *Matrix Inversion Lemma* can be used in this step to efficiently invert the matrix $\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$.

$$(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1} = \mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{\Psi}^{-1}$$

M-Step : Update $\mathbf{W}$ and $\mathbf{\Psi}$ as follows.

$$\mathbf{W}_{new} = (\textstyle\sum_{i=1}^{n}\mathbf{x}_i E(\mathbf{z}|\mathbf{x}_i)^T)(\sum_{j=1}^{n}E(\mathbf{z}\mathbf{z}^T|\mathbf{x}_j))^{-1}$$

$$\mathbf{\Psi}_{new} = \tfrac{1}{n}diag\{\textstyle\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T - \mathbf{W}_{new}E(\mathbf{z}|\mathbf{x}_i)\mathbf{x}_i^T)\}$$

Appendix B

Probability Computation in PWPCA Clustering

To implement PWPCA clustering, we rewrite the probability defined by equation (5.1) as follows:

$$
\begin{aligned}
p(\mathbf{x}) &\approx \frac{\exp\{-\frac{1}{2}\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i}\}}{(2\pi)^{q/2}\prod_{i=1}^{q}\lambda_i^{1/2}} \times \frac{\exp\{-\frac{1}{2\sigma}(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2)\}}{(2\pi\sigma)^{(p-q)/2}} \\
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i} - \ln\left((2\pi)^{q/2}\prod_{i=1}^{q}{\lambda_i}^{1/2}\right)\right\} \times \\
&\quad \exp\left\{-\frac{1}{2\sigma}\left(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2\right) - \ln(2\pi\sigma)^{(p-q)/2}\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i} + \sum_{i=1}^{q}\ln(\lambda_i)\right) - \frac{q}{2}\ln(2\pi)\right\} \times \\
&\quad \exp\left\{-\frac{1}{2\sigma}\left(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2\right) - \frac{p-q}{2}\ln\sigma - \frac{p-q}{2}\ln(2\pi)\right\}
\end{aligned}
$$

Adding a constant c to the exponent is the same as multiplying e^c to the exponential value, and therefore does not change the weights computed by normalizing $p(\mathbf{x})$'s across clusters. Since $\frac{q}{2}\ln(2\pi)$ and $\frac{p-q}{2}\ln(2\pi)$ are constants, those terms are dropped in the probability computation. Therefore,

$$
\begin{aligned}
p(\mathbf{x}) &\approx \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i} + \sum_{i=1}^{q}\ln(\lambda_i)\right)\right\} \times && \text{(B.1)} \\
&\quad \exp\left\{-\frac{1}{2\sigma}\left(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2\right) - \frac{p-q}{2}\ln\sigma\right\} && \text{(B.2)}
\end{aligned}
$$

The average trailing variance σ is different for each cluster. In practice, we observed that the outer-subspace probability is very sensitive to σ which makes the cluster size bouncing from iteration to iteration. For example, when p is order of hundred and q is order of ten, $\frac{p-q}{2}\ln\sigma$ can easily become order of thousand which then dominate the exponential term. It makes the outer-subspace probability extremely small so that it loses most of the samples in that iteration. In the next iteration, the σ gets small and the cluster attracts more samples than other clusters having larger σ. Although small σ makes the magnitude of the first term in the exponent increase, for the data set we tested, it does not have much influence on the magnitude of the entire exponent. Therefore, when σ values are radically different between clusters, the size of some clusters might change back and forth for each iteration.

After observed this phenomena, we made an assumption that all clusters have the same average trailing variance at each iteration. This is a rather strict assumption and might not be true in general. Although all samples are used for subspace computation in each cluster, samples are weighted first so those samples with negligible weights do not contribute to the computation. If q is not large enough to keep most variances within the subspace, different σ will be obtained for each cluster depending on the number of samples that had significant contribution to the subspace computation in the cluster. In our implementation, we set σ as the average of all σ's across clusters. Therefore, $\frac{p-q}{2}\ln\sigma$ can be dropped in equation (B.2) so the outer-subspace probability is solely decided by the first term in the exponent. In this case, σ determines the importance of the error $\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2$ in the probability computation. The effect of changing the magnitude of σ can be seen in Chapter 6.

Furthermore, to avoid numerical underflow caused by a very small negative exponent, we shift the exponents by some constant amount. Let $MaxExpInSpc$ and $MaxExpOutSpc$ be the maximum values of the exponents $(-\frac{1}{2}(\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i}+\sum_{i=1}^{q}\ln(\lambda_i))$ and $(-\frac{1}{2\sigma}(\|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{q} z_i^2))$ across clusters for a given sample $\mathbf{x}$, respectively. Then, the probability computed in the actual implementation is:

$$\begin{aligned} p(\mathbf{x}) \quad = \quad & \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{q}\frac{z_i^2}{\lambda_i}+\sum_{i=1}^{q}\ln(\lambda_i)\right)-MaxExpInSpc\right\}\times \\ & \exp\left\{-\frac{1}{2\sigma}\left(\|\tilde{\mathbf{x}}\|^2-\sum_{i=1}^{q}z_i^2\right)-MaxExpOutSpc\right\} \end{aligned}$$

Appendix C

Glossary

allocentric coordinates: a coordinate system centered on a point external to the viewer, whose body axes are not used to code location.

area 17: see *primary visual cortex.*

complex cell: the most common type of cells in the primary visual cortex that is sensitive to motion and has a relatively large receptive field. It is less sensitive to position than simple cells [114].

cones: a type of retinal photoreceptors that is active in daylight (photopic) conditions. The three different types of cone found in the human retina provide the basis for color vision [164].

dorsal pathway: one of two theorized systems of visual information processing. Information thought to progress toward the parietal cortex V1→V2→MT→PP (posterior parietal cortex). Functions in comprehension of spatial arrangement. See Figure 1.2.

egocentric coordinates: a coordinate system centered on the viewer.

ERP: see *event-related potential.*

event-related potential: the electromagnetic brain activity. The method of ERPs recording the brain activity such as EEG and MEG signals reflects real-time neuronal functioning.

fMRI: see *functional magnetic resonance imaging.*

functional magnetic resonance imaging: a technique used for imaging brain activity by visualizing changes of deoxygenated hemoglobin in the brain area that occur over timespans. The concentration of deoxygenated hemoglobin changes systematically where there is neural activity, and can be detected by the MRI scanner.

fusiform gyrus: the spindle-shaped convolutions on the surface of the cerebral hemispheres located in the medial part of the occipital lobe.

ganglion cells: the last neurons in the retina, whose axons exit the eye as the optic nerve. [114]. M-type ganglion cells are characterized by a large cell body, a transient response to light, and no sensitivity to different wavelengths of light. On the other hand, P-type ganglion cells are characterized by a small body, a sustained response to light, and sensitivity to different wavelengths of light [12].

Gaussian pyramid: in the Gaussian image pyramid, the resolution is decreased by successive convolutions of the image at the previous level of the pyramid with a Gaussianlike kernel. After the low-pass Gaussian convolutions, the sample density is typically decreased by sampling every other pixel [57].

image pyramid: a sequence of copies of an image in which both sample density and resolution are decreased in regular steps. The bottom level of the pyramid is the original image. Each successive level is obtained from the previous level by a filtering operator followed by a sampling operator [57].

infero-temporal (or inferior temporal) cortex: lower part of the temporal cortex.

inferior temporal lobe: lower part of the lobe of the brain that is inferior to the lateral sulcus and anterior to the occipital lobe; it is associated with auditory processing and olfaction.

Laplacian pyramid: in Laplacian image pyramid, each layer is obtained by taking the Laplacian of the corresponding level on the Gaussian pyramid. The Laplacian convolution kernel is typically defined as the kernel obtained by taking the Laplacian of a Gaussian having an appropriately chosen value for its standard deviation. It can be rapidly implemented by taking the difference between two successive layers in the Gaussian pyramid [57].

lateral geniculate nucleus: the thalamic nucleus that relays information from the retina to the primary visual cortex [12].

LGN: see *lateral geniculate nucleus.*

lateral occipital complex: the region located on the later bank of the fusiform gyrus extending ventrally and dorsally.

LOC: see *lateral occipital complex.*

magnocellular channel: the visual information processing channel that begins with the M-type retinal ganglion cells and leads to a layer of the primary visual cortex. It is believed to process information about visual movement [12].

middle temporal area: an area of the neocortex, at the junction of the parietal and temporal lobes, that receives input from the primary visual cortex and appears to be specialized for the detection of stimulus movement (also called V5) [12].

MT: see *middle temporal area.*

occipital lobe: the region of the cerebrum lying under the occipital bone [12]; the posterior lobe of the brain.

optic nerve: the bundle of ganglion cells that passes from the eye and carries visual information to the brain.

parvocellular channel: the visual information processing channel that begins with the P-type retinal ganglion cells and leads to a layer of the primary visual cortex. It is believed to process information about object shape [12].

PET: see *positron emission tomography.*

photoreceptors: the specialized nerve cells in the retina that transduce light energy into changes in membrane potential [12].

positron emission tomography: a technique used for imaging brain activity by measuring the flow of blood containing radioactive atoms that emit positrons [12].

posterial parietal cortex: the posterior region of the parietal lobe, located roughly 'after' vision and 'before' motor control in the cortical information processing hierarchy. It is involved in visual and somatosensory integration and attention [12].

posterial parietal lobe: the backside of the parietal lobe.

primary visual cortex: the first cortical visual area of the brain, which receives input directly from the LGN; also called area 17, striate cortex, and V1.

pulvinar: located in the posterior thalamus of the brain. The pulvinar receives input from the superior colliculus and projects to V1 [102].

receptive field: the region of the visual field in which a stimulus can evoke a change in the firing rate of the cell.

retinotopic: the topographic organization of visual pathways where the neurones from the retina are projected orderly so that the spatial structure of the image is preserved at the destination [164].

rods: a type of retinal photoreceptors that contain rhodopsin and, are specialized for low light levels [12].

repetitive transcranial magnetic stimulation: transcranial magnetic stimulation (TMS) utilizes an electromagnet placed on the scalp that generates magnetic field pulses roughly the strength of an MRI scan. The magnetic pulses pass readily through the skull and stimulate the underlying cerebral cortex. Low frequency (once per second) TMS has been shown to induce sustained reductions in cortical activation [168].

rTMS: see *repetitive transcranial magnetic stimulation.*

saccadic eye movements: the ballistic movements of the eyes that we use to explore the visual surroundings. The eyes jump from one fixation point in space to another by saccadic movements [12].

SC: see *superior colliculus.*

simple cell: the cells found in the primary visual cortex having an elongated orientation selective receptive field with distinct on and off subregions [12].

single-cell recording: a technique which records the action potential of a neuron without contamination ofaction potentials of neighboring neurons. For that, a small electrode is lowered into the brain area of interest while the subject is anesthetized. The electrode is positioned close to a cell whose action potential is to be recorded.

striate cortex: see *primary visual cortex.*

superior colliculus: a structure in the tectum of the midbrain that receives direct retinal input and directs saccadic eye movements [12].

thalamus: a paired structure of two tiny egg-shaped structures in the diencephalon. This structure is a crucial area for integrating and organizing sensory information that comes into the brain. In the thalamus, this information is processed and forwarded to the key cortical areas where more processing and integrating will take place [165].

V1: see *primary visual cortex.*

V2: a visual area which receives a somewhat patchy input from V1 and has a rather disorderly topographic organization. It is revealed that, in V2, a pattern of alternating thick and thin stripes, each separated by a thin interstripe region exists. The thick stripes are a part of the magnocellular channel, while the thin stripes and interstripes are a part of the parvocellular channel [167].

V3: a visual area which receives inputs from the thick stripes in V2, and from layer 4B in V1. Only the lower part of the visual field is represented in V3. Properties of cells in V3 offer few clues as to its function. Most are selective for orientation, and many are also tuned to motion and to depth. Relatively few are color sensitive [167].

V4: a visual area which receives input mainly from the thin and interstripe regions of V2. It also has connections from V1 and V3. Although the area contains many cells that are color selective, indicating a role in color analysis, cells are also found with complex spatial and orientation tuning, suggesting that the area is also important for spatial vision [167].

V5: see *middle temporal area.*

ventral pathway: one of two theorized systems of visual information processing. Information though to progress toward the temporal cortex V1→V2→V4→IT (inferior temporal cortex). Functions for analysis of object qualities such as pattern shapes, size and colors. See Figure 1.2.

visual cortex: the neocortex that appears to be directly involved in vision, with over twenty distinct areas. Some of the areas concerned are quite well understood, others are still a complete mystery.

visual mental imagery: a form of experience that resembles perceptual experience, but which occurs in the absence of the appropriate stimuli for the relevant perception; sometimes colloquially called visualization, or "seeing in the mind's eye" [166].

REFERENCES

[1] E.H. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt, and J.M. Ogden. Pyramid method in image processing. *RCA Engineer*, 29:33–41, 1984.

[2] E.H. Adelson, C.R. Carlson, and A.P. Pica. Modeling the human visual system. *RCA Engineer*, 27:56–64, 1982.

[3] E.H. Adelson and J.A. Movshon. Binocular disparity and the computation of two-dimensional motion. *Journal of the Optical Society of America*, 1(A):1266, 1984.

[4] C.H. Anderson and D.C. Van Essen. Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences, USA*, 84(17):6297–6301, 1987.

[5] K. Baek and B.A. Draper. Factor analysis for background suppression. In *International Conference on Pattern Recognition*, Québec City, Canada, August 2002, *To appear*.

[6] K. Baek, B.A. Draper, J.R. Beveridge, and K. She. PCA vs. ICA: A comparison on the FERET data set. In *The* 6^{th} *Joint Conference on Information Sciences*, pages 824–827, Durham, NC, March 2002.

[7] M.S. Bartlett. *Face Image Analysis by Unsupervised Learning.* Kluwer Academic, 2001.

[8] M.S. Bartlett, G. Donato, J.R. Movellan, J.C. Hager, P. Ekman, and T.J. Sejnowski. Image representations for facial expression coding. In S. Solla, T. Leen, and K. Mueller, editors, *Advances in Neural Information Processing Systems*, pages 886–892. MIT Press, Cambridge, MA, 2000.

[9] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, 1998.

[10] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Independent component representations for face recognition. *IEEE Transaction on Neural Networks*, 2002.

[11] A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley & Sons, Inc., New York, 1994.

[12] M.F. Bear, B.W. Connors, and M.A. Paradiso. *Neuroscience: Exploring the Brain*. Williams & Wilkins, 1996.

[13] J. Beck. On the computational modeling of human vision. In L. Davis, editor, *Foundations of Image Analysis*, chapter 1, pages 1–27. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

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

[15] A.J. Bell and T.J. Sejnowski. A non-linear information maximization algorithm that performs blind separation. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 467–474. MIT Press, Cambridge, MA, 1995.

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

[17] J.R. Bergen and B. Julesz. Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303:696–698, 1983.

[18] D. Besner. Visual pattern recognition: Size preprocessing re-examined. *Quarterly Journal of Experimental Psychology*, 35A:209–216, 1983.

[19] J.R. Beveridge and E.M. Riseman. Optimal geometric model matching under full 3D perspective. *CVGIP: Image Understanding*, pages 54–63, February 1994.

[20] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[21] I. Biederman. Visual object recognition. In S.F. Kosslyn and D.N. Osherson, editors, *An Invitation to Cognitive Science*, volume 2, chapter 4, pages 121–165. MIT Press, Visual Cognition, 1997.

[22] I. Biederman and M. Bar. One-shot viewpoint invariance in matching novel objects. *Vision Research*, 39:2885–2889, 1999.

[23] I. Biederman and P. Kalocsai. Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society: Biological Sciences*, 352:1203–1219, 1997.

[24] I. Biederman and M.M. Shiffrar. Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):640–645, 1987.

[25] J.A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, University of California at Berkeley, 1998.

[26] D. Bolme and B.A. Draper. Interpreting LOC and complex cell responses. Tübingen, Germany, 2002, *submitted.*

[27] M. Bressan, D. Guillamet, and J. Vitrià. Using an ica representation of high dimensional data for object recognition and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 1004–1009, December 2001.

[28] R.A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17:285–348, 1981.

[29] H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, USA*, 92:60–64, 1992.

[30] P.J. Burt and E.H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31:532–540, 1983.

[31] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, 1997.

[32] K.R. Cave and S.M. Kosslyn. Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, 118(2):148–164, 1989.

[33] V.P. Clark, K. Keil, J.M. Maisog, S. Courtney, L.G. Ungeleider, and J.V. Haxby. Functional magnetic resonance imaging of human visual cortex during face matching: A comparison with positron emission tomography. *Neuroimage*, 4:1–15, 1996.

[34] P. Comon. Independent component analysis - a new concept? *Signal Processing: Special Issue on High-Order Statistics*, 36(3):287–314, 1994.

[35] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–39, 1977.

[36] R. Desimone and L.G. Ungerleider. Neural mechanisms of visual processing in monkeys. In F. Boller and J. Grafman, editors, *Handbook of Neuropsychology*, pages 267–299. Elsevier, Amsterdam, 1989.

[37] S. Dickinson. Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15(4):277–292, 1997.

[38] S. Dickinson, I. Biederman, A. Pentland, J.-O. Eklundh, R. Bergevin, and R. Munck-Fairwood. The use of geons for generic 3-d object recognition. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1693–1699, Chambery, France, August 1993.

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

[40] C.J. Downing and S. Pinker. The spatial structure of visual attention. In M.I. Posner and O.S.M. Marin, editors, *Attention and Performance XI: Mechanics of Attention*, pages 171–187. Erlbaum, Hillsdale, NJ, 1985.

[41] B.A. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding: Special Issue on Face Recognition*, submitted.

[42] R.O. Duda, P.E. Hart, and Stork D.G. *Pattern Classification*. John Wiley and Sons, Inc., New York, 2001.

[43] S. Edelman and H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400, 1992.

[44] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Palo Alto, CA: Consulting Psychologists Press, 1978.

[45] C.W. Eriksen and J.E. Hoffman. Selective attention: Noise suppression or signal enhancement? *Bulletin of the Psychonomic Society*, 4:587–589, 1974.

[46] P.T. Fox, F.M. Miezin, J.M. Allman, D.C. Van Essen, and M.E. Raichle. Petinotopic organization of human visual cortex mapped with positron-emission tomography. *Journal of Neuroscience*, 7(3):913–922, 1987.

[47] B.J. Frey, A. Colmenarez, and T.S. Huang. Mixtures of local linear subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[48] I. Fujita, M.I. Tanaka, and K. Cheng. Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346, 1992.

[49] I. Gauthier and M.J. Tarr. Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682, 1997.

[50] I. Gauthier and M.J. Tarr. Orientation priming of novel shapes in the context of viewpoint dependent recognition. *Perception*, 26:51–73, 1997.

[51] I. Gauthier and M.J. Tarr. Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior (in press). *JEP:HPP*, 2001.

[52] I. Gauthier, M.J. Tarr, A.W. Anderson, P. Skudlarski, and J.C. Gore. Behavioral and neural changes following expertise training. In *presented at Annual Meeting of the Psychonomic Society*, Philadelphia, PA, 1997.

[53] I. Gauthier, M.J. Tarr, A.W. Anderson, P. Skudlarski, and J.C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Neuroscience*, 2(6):568–573, 1999.

[54] I. Gauthier, M.J. Tarr, J. Moylan, A.W. Anderson, P. Skudlarski, and J.C Gore. Does visual subordinate-level categorization engage the functionally defined fusiform face area? *Cognitive Neuropsychology*, 17:143–163, 2000.

[55] Z. Ghahramani and G.E. Hinton. The em algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, 1997.

[56] W.E.L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, Cambridge, MA, 1990.

[57] R.M. Haralick and L.G. Shapiro. *Computer and Robot Vision*. Addison Wesley, 1993.

[58] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[59] J.V. Haxby, L.G. Ungerleider, V.P. Clark, J.L. Schouten, E.A. Hoffman, and A. Martin. The effect of face inversion on activity in human neural systems for face and object recognition. *Neuron*, 22:189–199, 1999.

[60] W.G. Hayward and M.J. Tarr. Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23:1511–1521, 1997.

[61] W.G. Hayward and P. Williams. Viewpoint dependence and object discriminability. *Psychological Science*, 11(1):7–12, 2000.

[62] S.H.C. Hendry and T. Yoshioka. A neurochemically distinct third channel in the macaque dorsal lateral geniculate nucleus. *Science*, 264:575–577, 1994.

[63] G.E. Hinton, M. Revow, and P. Dayan. Recognizing handwritten ditits using mixtures of linear models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 1015–1022. MIT Press, 1995.

[64] D.H. Hubel and T.N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

[65] J.E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517, 1992.

[66] D.P. Huttenlocher and S. Ullman. Recognizing solid object by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[67] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10:1–5, 1999.

[68] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, New York, 2001.

[69] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[70] A. Ishai, L.G. Ungerleider, A. Martin, and J.V. Haxby. The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, 12:35–51, 2000.

[71] A. Ishai, L.G. Ungerleider, A. Martin, J.L. Schouten, and J.V. Haxby. Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, USA*, 96:9379–9384, 1999.

[72] R.A. Jacobs and S.M. Kosslyn. Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representation. *Cognitive Science*, 18:361–386, 1994.

[73] C.-S. Kalanit, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41:1409–1421, 2001.

[74] P. Kalocsai and I. Biederman. Differences of face and object recognition in utilizing early visual information. In W. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, 1998.

[75] N. Kambhatla and T.K. Leen. Dimension reduction by local PCA. *Neural Computation*, 9:1493–1516, 1997.

[76] N. Kanwisher. Domain specificity in face perception. *Neuroscience*, 3(8):759–763, 2000.

[77] N. Kanwisher, M. Chun, J. McDermott, and P. Ledden. Functional imaging of human visual recognition. *Cognitive Brain Research*, 5:55–67, 1996.

[78] N. Kanwisher, J. McDermott, and M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17:4302–4311, 1997.

[79] M. Kirby and L. Sirovich. Applications of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[80] J.J. Koenderink and A.J. Van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[81] A. Koriat, J. Norman, and R. Kimchi. Recognition of rotated letters: Extracting invariance across successive and simultaneous stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, 17:444–457, 1991.

[82] S.M. Kosslyn. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA, 1994.

[83] S.M. Kosslyn. Visual mental images and re-presentations of the world: A cognitive neuroscience approach. In J.S. Gero and B. Tversky, editors, *Visual and Spatial Reasoning in Design*. MIT, Cambridge, MA, 1999.

[84] S.M. Kosslyn, N.M. Alpert, W.L. Thompson, C.F. Chabris, S.L. Rauch, and A.K. Anderson. Identifying objects seen from different viewpoints: A PET investigation. *Brain*, 117:1055–1071, 1994.

[85] S.M. Kosslyn, A. Pascual-Leone, O. Felician, J.P. Keenan, W.L. Thompson, G. Ganis, K.E. Sukel, and N.M. Alpert. The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science*, 284:167–170, 1999.

[86] S.M. Kosslyn, J. Shephard, and W.L. Thompson. The neurofunctional organization of late visual processing. In B. Mazoyer, editor, *Tutorials in Cognitive Neuroscience*. Academic Press, New York, In press.

[87] Z. Kourtzi and N. Kanwisher. Cortical regions involved in perceiving object shape. *The Journal of Neuroscience*, 20(9):3310–3318, 2000.

[88] P. Kruizinga, N. Petkov, and S.E. Grigorescu. Comparison of texture features based on gabor filters. *Proceedings of the 10th International Conference on Image Analysis and Processing*, pages 142–147, September 1999.

[89] T.S. Lee, D. Mumford, R. Romero, and V.A.F. Lamme. The role of the primary visual cortex in higher lever vision. *Vision Research*, 38:2429–2454, 1998.

[90] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *International Conference on Audio and Video Based Biometric Person Authentication*, Washington, D.C., 1999.

[91] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, Boston, 1985.

[92] D.G. Lowe. Three-dimensional object recognition from single two-dimensional image. *Artificial Intelligence*, 31:355–395, 1987.

[93] D.G. Lowe. The viewpoint consistency of computer vision. *International Journal of Computer Vision*, 1:57–72, 1987.

[94] D.G. Lowe. Fitting parameterized 3-d models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.

[95] E. Maguire, C.D. Frith, and L. Cipolotti. Distinct neural systems for the encoding and recognition of topography and faces. *NeuroImage*, 13:743–750, 2001.

[96] D. Marr. *Vision*. Freeman, Cambridge, MA, 1982.

[97] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London*, 290(B):199–218, 1980.

[98] C.J. Marsolek. *Visual Form Systems in the Cerebral Hemispheres*. PhD thesis, Harvard University, 1992.

[99] G.G. Matthews. *Neurobiology: Molecules, Cells, and Systems*. Blackwell Science, MA, 2001.

[100] B.W. Mel. SEEMORE: A view-based approach to 3D object recognition using multiple visual cues. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Systems 8*, pages 865–871. MIT Press, 1996.

[101] P. Michelon and O. Koenig. Imagery feedback in object identification. In *Object Perception and Memory Conference Proceedings*, pages 69–70, 1999.

[102] A.D. Milner and M.A. Goodale. *The Visual Brain in Action*. Oxford University Press, Oxford, 1995.

[103] B. Milner. Visual recognition and recall after right temporal-lobe excision in man. *Neuropsychologia*, 6:191–209, 1968.

[104] M. Mishkin, L.G. Ungerleider, and K.A. Macko. Object vision and spatial vision: Two cortical pathways. *Trends in Neuroscience*, 6:414–417, 1983.

[105] B. Moghaddam. Principal manifolds and bayesian subspaces for visual recognition. In *International Conference on Computer Vision*, Corfu, Greece, 1999.

[106] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Pattern Aanlysis and Machine Intelligence*, 19(7):696–710, 1997.

[107] H. Moon and J. Phillips. Analysis of PCA-based face recognition algorithms. In K. Boyer and J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, Los Alamitos, CA, 1998.

[108] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[109] K. Nakamura, R. Kawashima, N. Sato, A. Nakamura, M. Sugiura, T. Kato, K. Hatano, K. Ito, H. Fukuda, T. Schormann, and K. Zilles. Functional delineation of the human occipito-temporal areas related to face and scene processing: A PET study. *Brain*, 123:1903–1912, 2000.

[110] S.K. Nayar, S. Nene, and H. Murase. Real-time 100 object recognition system. In *Proceedings of ARPA Image Understanding Workshop*, San Francisco, CA, February 1996.

[111] R. Nelson and A. Selinger. Large-scale tests of a keyed, appearance-based 3D object recognition system. *Vision Research, Special issue on Computational Vision*, 38(15–16), August 1998.

[112] K.M. O'Craven and N. Kanwisher. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, 12:1013–1023, 2000.

[113] B.A. Olshausen, C.H. Anderson, and D.C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719, 1993.

[114] S.E. Palmer. *Vision Science: Photons to Phenomenology.* MIT Press, 1999.

[115] A.P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986.

[116] N. Petkov and P. Kruizinga. Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented stimuli: Bar and grating cells. *Biological Cybernetics*, 76:83–96, 1997.

[117] Z. Pizlo. A theory of shape constancy based on perspective invariants. *Vision Research*, 34(12):1637–1658, 1994.

[118] M.D. Plumbley. A network which performs orthonormalized principal subspace extraction. Technical Report 94/06, King's College London, 1994.

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

[120] T. Poggio and A. Hurlbert. Observations on cortical mechanisms for object recognition and learning. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 153–182. MIT Press, 1994.

[121] D.A. Pollen, J.P. Gaska, and L.D. Jacobson. Physiological constraints on models of visual cortical function. In M. Rodney and J. Cotterill, editors, *Models of Brain Functions*, pages 115–135. Cambridge University Press, New York, 1989.

[122] C.Y. Pong and J.H. Lai. Independent component analysis of face images. In S.-W. Lee, H. Bülthoff, and T. Poggio, editors, *Lecture Note in Computer Science No. 1811: International Workshop on Biologically Motivated Computer Vision*, pages 545–553. Springer-Verlag, 2000.

[123] M.I. Posner, C.R.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174, 1980.

[124] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientic Computing.* Cambridge University Press, New York, 1992.

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

[126] R.P.N. Rao and D. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505, 1994.

[127] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[128] D. Rubin and D. Thayer. Em algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[129] D. Rubin and D. Thayer. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2):115–125, 1999.

[130] J. Rubner and P. Tavan. A self-organizing network for principal component analysis. *Europhysics Letters*, 10:693–698, 1989.

[131] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. Technical Report 453, M.I.T. Media Laboratory, Perceptual Computing Section, December 1997.

[132] C. Schmid and R. Mohr. Matching by local invariants. Technical Report 2644, INRIA, August 1995.

[133] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96)*, San Francisco, CA, June 1996.

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

[135] R.N. Shepherd and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[136] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[137] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[138] D. Swets and J. Weng. Hierarchical discriminant analysis for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):386–401, 1999.

[139] B.G. Tabachnick and L.S. Fidell. *Using Multivariate Statistics*. Allyn & Bacon, Inc., Boston, 2000.

[140] J.W. Tanaka and T. Curran. A neural bases for expert object recognition. *Psychological Science*, 12:43–47, 2001.

[141] J.W. Tanaka and M.J. Farah. Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, 46A:225–245, 1993.

[142] J.W. Tanaka and I. Gauthier. Expertise in object and face recognition. In R. Goldstone, P. Schyns, and D.L. Medin, editors, *Psychology of Learning and Motivation*, pages 83–125. Academic Press, San Diego, CA, 1997.

[143] M.J Tarr. Rotating objects to recognize them: A case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2:55–82, 1995.

[144] M.J Tarr. Visual object recognition: Can a single mechanism suffice? *Essay submitted to the James S. McDonnell Centennial Fellowship Competition*, 1997.

[145] M.J. Tarr and I. Gauthier. FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, 3(8):764–769, 2000.

[146] M.J. Tarr, P. Williams, W.G. Hayward, and I. Gauthier. Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, 1(4):275–277, 1998.

[147] M.L. Teixeira and B.A. Draper. Imagery as a biologically motivated enhancement to PCA-based face matching. Tübingen, Germany, 2002, *submitted*.

[148] M. Tipping and C. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[149] F. Tong, K. Nakayama, M. Moscovitch, O. Weinrib, and Kanwisher N. Response properties of the human fusiform face area. *Cognitive Neuropsychology*, 17:257–279, 2000.

[150] R.B.H. Tootell, M.S. Silverman, E. Switkes, and De Valois R.L. Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, 218:902–904, 1982.

[151] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[152] S Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32:193–254, 1989.

[153] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005, 1991.

[154] D.C. Van Essen, H.A. Drury, S. Joshi, and I. Miller. Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces. *Proceedings of the National Academy of Science, USA*, 95:788–795, February 1998.

[155] M. Vitkovitch and G. Underwood. Hemispheric differences in the processing of pictures of typical and atypical semantic category members. *Cortex*, 27:495–480, 1991.

[156] P. Williams. *Prototypes, Exemplars, and Object Recognition*. PhD thesis, Brown University, 1997.

[157] P. Williams, I. Gauthier, and M.J. Tarr. Feature learning during the acquisition of perceptual expertise. *Behavioral and Brain Sciences*, 21(1):40–41, 1998.

[158] W. Yambor, B.A. Draper, and J.R. Beveridge. Analysis of PCA-based face recognition algorithms: Eigenvector selection and distance measures. In *Second Workshop on Empirical Evaluation Methods in Computer Vision*, Dublin, July 2000.

[159] P.C. Yuen and J.H. Lai. Independent component analysis for face images. In *IEEE Workshop on Biologically Motivated Computer Vision*, pages 545–553, Seoul, Korea, 2000.

[160] S.M. Zeki, Watson J.D.G., C.J. Lueck, Friston K.J., Kennard C., and Frackowiak R.S.J. A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, 11:641–649, 1991.

[161] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J Weng. Discriminant analysis of principle components for face recognition. In *Third International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 1998.

[162] http://www.dur.ac.uk/˜dps0ta/visual_system_II.pdf.

[163] http://www.cs.rug.nl/˜imaging/simplecell.html.

[164] http://www.blackwellpublishers.co.uk/psychol/glossary.htm.

[165] http://www.childtrauma.org/brain_I.htm.

[166] http://plato.stanford.edu/entries/mental-imagery/.

[167] http://www.biols.susx.ac.uk/home/George_Mather/Linked%20Pages/Physiol/Cortex.html.

[168] http://info.med.yale.edu/psych/clinics/rtms.html.

Index