Biologically Inspired Visual Models by Sparse and Unsupervised Learning

Li Yang

M.S. in EE, University of Electronic Science & Technology of China, 2001

B.S. in Electrical Engineering, Yunnan University, 1992

B.A. in Chinese Literature, Yunnan University, 1992

A dissertation submitted to the faculty of the
OGI School of Science & Engineering
at Oregon Health & Sciences University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical Engineering and Computer Science

January 2007

The dissertation “Biologically Inspired Visual Models by Sparse and Unsupervised Learning” by Li Yang has been examined and approved by the following Examination Committee:

Misha Pavel, Ph.D.
Professor
OGI School of Science & Engineering
Thesis Research Adviser

Xubo Song, Ph.D.
Assistant Professor
OGI School of Science & Engineering

Dan Hammerstrom, Ph.D.
Professor
Portland State University

Dedication

To my parents, for their love and support always.

Acknowledgements

My deepest thanks go to my parents, to whom I owe so much, but have always returned so little. I am short of words to express my thanks to my beloved husband and daughter for their love and support throughout this endeavor.

I would also like to thank the members of my committee for their efforts and time spent on my behalf. I greatly thank my adviser, Dr. Misha Pavel, for accepting me as his student. It is my honor to be his student. I have benefited enormously from his breadth of knowledge and capacity for insight for the past years. It has been a great privilege to work with Dr. Dan Hammerstrom. He edited my dissertation word by word, sentence by sentence, giving lots of corrections on both English and technique aspects. Dr Xubo Song has continuously given suggestions and encouragements on my Ph.D. study since I came to OGI in 2001. At the end, she still took time to review my dissertation and gave valuable suggestions on modifying it. I also want to thank Dr. Marwan Jabri for his guidance on my Ph.D. study.

Finally, I would especially thank Ms. Lisa Pickert, Ms. Penny Waldrep and Dean Edward Thompson. Without their kindly help and encouragements, I would never complete this work.

This work was supported by NASA Revolutionary Computing Program under contract NCC-2-1253.

Table of Contents

Table of Figures

Abstract

Biologically Inspired Visual Models by Sparse and Unsupervised Learning

Li Yang
Ph.D., Department of Computer Science and Electrical Engineering
OGI School of Science & Engineering at OHSU, 2006

Advisor: Misha Pavel

The research objective was to develop visual models and corresponding algorithms to automatically extract parts of novel objects from raw gray-scale images, and finally, to achieve object recognition with rotational invariance within structural description framework. Given primate visual systems can process visual information fast and well, we used primate brains as the inspiration for developing visual models with sparse and unsupervised learning algorithms. Sparse representation allowed our visual models to exhibit intrinsic fault-tolerance and low-power consumption operation compared to other computing paradigms. Unsupervised learning allowed our visual models to automatically extract features of novel objects based on statistical properties of input images, and without the visual models employing any explicit knowledge.

Inspired by the primate visual ventral pathway, we developed several visual models in hierarchical network architecture for low-level visual feature extraction (V1 and V2 models), parts-based shape representation (V4 model) and high-level object recognition (IT model). Using this world as its own representation and extracting information from it, as necessary, through the action of feature detectors based on the notion of cells' receptive fields, our visual models are biologically inspired and are also computationally tractable.

Our results show that these models can efficiently and adaptively process visual information with approximate transformation invariance. The low-level features extracted by the V1 and V2 models are very sparse but rich enough for further visual processing in high-layer models, such as V4 and IT. With the sparse coding constraint, the V4 model combines unsupervised representation in the feed-forward stream with lateral interaction to achieve stable, efficient and natural representation of shapes. Furthermore, we found that V4 model cells display same curvature and object centered tuning as the reported tuning properties of V4 cells in the primate visual ventral pathway. Based on object parts output from the V4 model, we developed an IT mode for the purpose of recognizing objects from different viewing angles, where objects are represented as flexible constellations of rigid parts. The IT model achieves very good object recognition results with approximate viewpoint invariance.

The main contribution of this work is the biologically motivated integration of a number of existing approaches, e.g., unsupervised learning and sparse representation into the hierarchical network architecture. These models yield better performance than many existing algorithms and represent biologically plausible mechanisms, therefore, may provide some idea to further explore the mechanisms of visual information processing both in biological and robotic settings.

Chapter 1

Introduction

1.1 Invariant Object Recognition

The human visual system is a versatile mechanism that enables us to interpret energy stimulating our eyes with remarkable sensitivity and sophistication. It enables us to process a stream of images and to recognize a multitude of familiar and novel objects as well as their relationships. The visual system does this with little effort, despite the fact that these objects may vary somewhat in form, color, texture, etc. Objects are recognized from many different vantage points (from the front, side, or back), in many different places, and in different sizes. One of the most fundamental and essential properties of the visual system is the ability to recognize a particular object, despite great variations in the images that impose on the retina (Giboson, Kaplan et al. 1969). Objects can even be recognized when they are partially obstructed from view (Gibson 1950). While it may be obvious that people are capable of recognizing objects under many variations in conditions, it is very difficult for object recognition in machine vision to be largely invariant with regard to changes in the size, position, and viewpoint of the object. Especially, invariance in viewpoint (invariance with respect to rigid transformations) provides the greatest challenge to object recognition theory (Biederman and Cooper 1991).

Visual information falling on the retina when a particular object is viewed varies drastically from occasion to occasion, depending on the distance from the image (which affects the size of the image on the retina), the vantage point from which the object is viewed, and the location of the object relative to the viewer (which affects the part of the retina that is stimulated). Therefore, the capability of recognizing objects with size, position and rotation invariance is very crucial to any visual system.

Size (or scale) Invariance: Objects can be recognized despite variations in actual or apparent size. When a particular object impinges on the retina, the size of the retinal image often varies from occasion to occasion. The primary cause of retinal image size variations is viewing distance. When an object is viewed from farther away, the retinal image will undoubtedly be smaller than when the object is viewed from a closer distance. If you hold your hand just a few inches from your face, for example, it fills almost your entire visual field, but if you move it to arm's length, it covers a much smaller area. The visual system must therefore disentangle the effects of object size and object distance in interpreting retinal size. Size constancy (or invariance) is the ability of the visual system to infer that a given object is the same size when it is viewed from different distances. Size invariance allows for the recognition of an object, even though the retinal surface area that the image activates is quite different (Biederman and Cooper 1992; Fiser and Biederman 1995). Therefore, it is possible to recognize both of the objects in Fig. 1 as the same parakeet, even though one of the objects is half the size of the other (Kirkpatrick 2001). A related phenomenon to size invariance is the ability to use the size of a retinal image to infer perception of distance so that objects that are smaller are inferred as being farther away. In a 3-D environment this inference would operate correctly, but in a 2-D image, the inference can create illusions: although the smaller parakeet is actually above

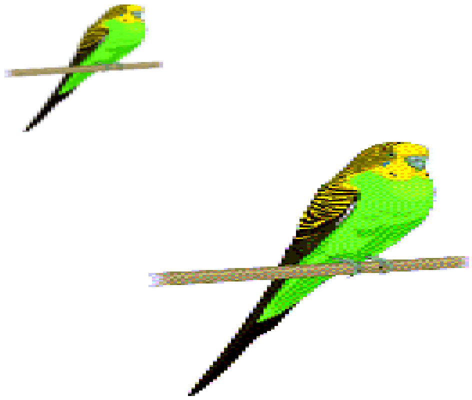


Figure 1.1 Size invariance: we recognize both of the above objects as the same parakeet, even though one of the objects is half the size of the other (Kirkpatrick 2001).

and to the left-of the larger parakeet, the visual system interprets the smaller bird as being in the distance. Combined, these two phenomena compose the size-distance paradox, an ability that allows us, as humans, to recognize an object from different distances and to use the size of an object to aid in judging the object's distance.

Position Invariance: When an object is moved to a new position in the environment, a different portion of the retina is stimulated. Nonetheless, modest changes in position do not disrupt recognition accuracy in human subjects; that is, object recognition is positional invariant. A crucial global property of real-world objects is their position in the environment. There are at least two different ways in which objects' position can be perceived: relative to observer's body and relative to various other objects in the environment. When an object is viewed, its position on the retina likely changes from occasion to occasion. It seems almost absurd to claim that the visual system's ability to perceive unmoving objects as stationary is an accomplishment, but when you think about it, it is. We typically move our eyes dozens of times each minute, and every one of these eye movements causes the images of all stationary objects to a new position (Ungerleider and Mishkin 1982). For example, every rightward movement of the eye causes the images of all stationary objects to shift leftward relative to the visual field. Yet in spite of their zigzagging retinal trajectories stationary objects appear rock-solid in their environmental position.

In fact, even within a single viewing, the retinal position that the image activates likely changes with movements of the observer's head and body. For the sake of processing efficiency alone, it seems plausible that the visual system might independently code the shape and position of an object, so that a representation of object shape would not have to be duplicated for every retinal position that is encountered. If such dual representations were to exist, it would allow for effortless identification of object shape attributes, despite variations in position. As a result, complete positional invariance would be achieved. In humans, complete position invariance has been reported without any measurable costs in recognition accuracy or speed (Ellis, Allport et al. 1989; Biederman and Cooper 1991).

Rotation Invariance: The major challenge to object recognition occurs when the position of the observer changes so that he or she views the object from a different perspective. In most case, this change in viewing geometry causes a transformation in the shape of its projected image. Object recognition in humans appears to occur despite differences in viewpoint (Biederman and Gerhardstein 1993; Biederman and Bar 1999), but there has been considerable controversy regarding whether true rotational invariance occurs (Biederman and Gerhardstein 1995; Tarr and Bülthoff 1995). A growing literature has demonstrated that there are costs in recognition accuracy or speed that are directly related to the degree of rotation from the familiar viewing orientation (Shepard and Metzler 1972; Bartram 1974; Hayward and Tarr 1997; Tarr, Williams et al. 1998). These controversies are primarily relevant in discriminating between rival theoretical accounts of object recognition. Regardless of the exact mechanism by which novel viewpoints are recognized, it is clear that humans do spontaneously recognize, to varying degrees, objects when seen from novel vantage points. The question is how the visual system is able to recognize objects with rotational invariance.

There are several possibilities for object recognition with rotational invariance (Palmer 1977; Rock and DiVita 1987; Palmer 1989; Palmer 1999). One is that our intuition about object recognition is based primarily on situations in which we perceive the same object from different vantage points by continuously moving from one view to another. Under these conditions, the perspective that changes from moment to moment are quite gradual and easily perceived as shape preserving, especially when an object is being viewed continuously. Another possibility is that much of the time our perception of an object's shape is closely correlated with its identity. That is, we may believe that object recognition with rotational invariance occurs simply because we are able to recognize the same object from different perspective views by using different features. For example, you may be able to recognize a friend's face both from a front view and from a profile view of his or her head. Even so, we might not be nearly as good at performing the same task on two different perspective views of a complete stranger's face or of some novel object. The third possibility for why object recognition with rotational invariance appears to be good in everyday situations is that most objects we see have features that distinguish some particular parts of them as the "front" and "back".

We note that the invariance with respect to the set of rigid transformations is only an approximation of the characteristics of the primate visual system. While there are exceptions to the invariance principles, such as the "face recognition" and the retinal eccentricity effects, our understanding of the mechanisms underlying the performance of the human visual system is critical in building systems that aim to achieve human performance.

1.2 Theories of Invariant Object Recognition

The problem in object recognition is to determine which, if any, of a given set of objects appear in a given image or image sequence (Tarr and Bulthoff 1998). Thus object recognition is a problem of matching models from a database with representations of those models extracted from raw images. Two main computational theories have been proposed to address invariant object recognition: structural description theory (Marr 1982; Biederman 1987; Mel 1997; Riesenhuber and Poggio 1999) and viewpoint-specific theory (Koenderink and van Doorn 1979; Huttenlocher and Ullman 1987; Ullman and Basri 1991).

Structural Description Theories: Perhaps the single most influential theoretical approach to object recognition over the past several decades is an extension of ideas behind structural description theories. Versions of it have been developed by a number of computer scientists and computationally oriented psychologists (Binford 1971; Marr and Nishihara 1978; Biederman 1987; Hummel and Biederman 1992; Mel 1997; Weber, Welling et al. 2000; Shams and Malsburg 2002). Of the many specific theories that have been advanced within the structural description framework, we will consider one in detail: Irving Biederman's (1987, 1992) recognition by components (RBC) theory, especially considering that RBC theory has been a foundation of object recognition since it was presented in 1987. The RBC theory is not radically different from several others but is somewhat easier to describe and has been developed with more attention to the results of experimental evidence. It is presented as representative of the structural description theories rather than as the “correct” one or even the “best” one.

Derived from the work of David Marr (Marr 1982), Irving Biederman’s RBC (1987) is a theory of object recognition in humans that accounts for the successful identification of objects despite changes in the size or rotation of the object. It is based on the idea that objects can be specified as spatial arrangements of primitive volumetric components (i.e. geons or parts). Moreover, RBC explains how moderately occluded or degraded images, as well as novel examples of objects, are successfully recognized by the visual system. The major contribution of RBC is the proposal that the visual system extracts geons (or geometric ions) and uses them to recognize objects.

RBC proposes that representations of objects are stored in the brain as structural descriptions. A structural description contains a specification of the object’s geons and their interrelations (e.g., a cube is above a cylinder). This means that RBC representations are essentially hierarchical networks the nodes of which correspond to 3-D geons or parts and whose links correspond to relations among these geons or parts. A perceived object is analyzed by the visual system, which parses the object into its constituent geons and parts. Then, the interrelations are determined, which include aspects such as relative position and size (e.g., the lamp shade is left-of, below, and larger-than the fixture). The geons and interrelations of the perceived object are matched against stored structural descriptions. If a reasonably good match is found, then successful object recognition will occur.

In RBC, geons are generalized cylinders whose inherently continuous parameters have been divided into a few discrete ranges that are fairly easy to distinguish from most vantage points. This produces a relatively small set of distinct primitive volumes from which a huge number of object representations can be constructed by putting two or more together, much as letters serve as a small set of primitives from which an enormous number of words can be constructed. Biederman defined a set of 36 qualitatively different geons and further constructed 108 quantitatively geons from these 36 geons with three aspect ratios. The ratio of the length of sweeping axis to that of the largest dimension of the cross-sectional area is called the aspect ratio. Because complex objects are conceived in RBC as structural descriptions that specify both the geons present and their spatial relations. If geons are the alphabet of complex objects, then spatial relations among geons are analogous to the order of letters in words. Just as it is possible to construct different words by arranging the same letters differently (e.g., BAT versus TAB), so is it possible to construct different object types by arranging the same geons differently (see Fig. 2).

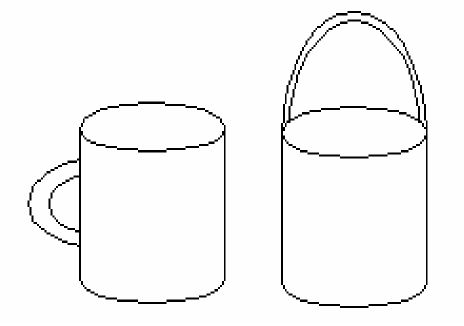


Figure 1.2 The cup and the pail are comprised of two components: a cylinder and a curved handle. However, the orientation and position of the handle relative to the cylinder differs between the objects. In order to discriminate the cup from the pail, one must be able to recognize the differences in the organization of the components, a more global property of objects.

To represent the structure of complex objects consisting of two or more geons, RBC uses structural descriptions in which 108 qualitatively different relations can be represented between two geons. Some of these relations concern how they are attached (e.g., side-connected and top-connected); others concern their relational properties, such as relative size (e.g., larger-than and smaller-than). With these 108 geon relations and 108 geons, it is logically possible to construct more than a million different two-geon objects.

Once the shape of an object has been represented via its component geons and their spatial relations, the problem of object recognition within RBC theory reduces to the process of matching the structural description of an incoming object with the set of structural description entry-level categories. Biederman has estimated that most people know about 30,000 different object categories, indicating that there are far more possible combinations of geons than there are actual objects. This is important if the visual system is to be able to realize that a novel object is not a member of any known category, for there must be arrangement of geons that do not correspond to known categories.

The actual mechanisms of RBC theory were not fully specified in the original formulation of RBC (Biederman 1987). They were supposed to be accomplished by spreading activation within a complex neural network but actual models of RBC were not specified. Since then, John Hummel and Biederman (1992) have devised a neural network implementation of RBC (called JIM, for "John and Irv's Model"). One of the key proposals in JIM is that many different attributes can be temporarily bound to a given geon (i.e. associated with that geon) by their units firing in synchrony. This mechanism, which is called dynamic binding, is an extremely useful and economical way to represent the association of different attributes that are part of the same component. It is also the mechanism that JIM uses to parse objects into geons. Different attributes are bound to different geons in the same representation by virtue of there being several different temporal patterns of activation present at the same time, one set of attribute units firing synchronously in one pattern of activation and another set of attributes firing synchronously in another pattern. These patterns of activation could be as simple as different temporal firing rates, provided the difference was large enough.

Although many details of JIM's operation are well specified, some had not yet been worked out at the time Hummel and Biederman wrote their 1992 article. The model has since been fleshed out more fully and is capable of recognizing many simple objects at an entry level correctly from a clean line drawing (Hummel 1999; Hummel 2000). The model is not yet able to start with raw gray-scale images, but this is true of most theories of high-level object recognition.

For all of the virtues of RBC as a theory of invariant object recognition, it has some problems as well. One of the most serious was the lack of representational power. Some of this comes from limitations of generalized cylinders as representational primitives for shape information. But these problems are aggravated by discretely classifying generalized cylinders into just 108 geons and reducing all possible geon relations to just 108 binary relations, because these restrictions further reduce the amount of visual detail geon descriptions can encode. Even the trillions of logically possible combinations are not sufficient to capture the subtle differences in shape that people normally and regularly detect in recognizing (classifying and identifying) objects. Another potential problem with RBC as a theory of object recognition is that it is not yet clear that RBC will actually be able to recognize objects from images, much less perform as humans do. The proof must await a fully implemented computer simulation that is actually able to identify common objects from gray-scale images as people obviously can do. RBC sounds quite plausible as a conceptual theoretical framework in many respects, but one of the most important lessons that have been learned from research in computer vision is that even plausible theories often do not work as expected or do so only under very restricted conditions. Hummel and Biederman's (1992) neural network implementation of RBC theory avoids many difficult problems by starting with hand-coded feature descriptions rather than gray-scale images. It thus skips what has turned out to be one of the most difficult problems with theories based on volumetric structural descriptions: identifying or extracting the primitive volumetric components from gray-scale information.

View-Specific Theories: Palmer, Rosch, and Chase (Palmer, Rosch et al. 1981) systematically investigated and documented perspective effects in object recognition. They began by having subjects view many of pictures of the same object and make subjective ratings of how much each one looked like the objects they depict using a scale from 1 (very much like) to 7 (very unlike). Then, they had other subjects name the categories of the objects shown, as quickly as possible, using several different perspectives for each of the objects. Their results showed that the pictures that were rated as the best views were named fastest and that naming latencies gradually increased as the "goodness" of the views declined, the ones rated "worst" being named much more slowly than the "best". Palmer et al. called the best, most easily identified view for each objects its canonical perspective.

The existence of canonical perspective and perspective effects in general shows the error in the intuition that we can recognize objects equally well from all possible perspective views and raises the possibility that objects might be identified by matching 2-D input views directly to some kind of view-specific category representation. To begin with the most obvious possibility, the idea of canonical perspective suggests a particularly simple view-specific, template-matching theory: each object might be represented by in memory by a single, canonical 2-D view, and 2-D input views might be matched directly against it. Canonical views would then be quickly and accurately categorized because they could be matched directly to the canonical 2-D representation in memory. Other input views would not match initially but would have to be rotated into alignment with the canonical view to achieve a successful 2-D match. Alternatively, the canonical perspective view in memory might be rotated into alignment with the input image. Performing such transformations would require more time and would increase the possibility of making errors, consistent with the reported pattern of perspective effects (Palmer 1983; Edelman and Bülthoff 1992).

However, there are at least three significant problems for such single-view template matching theory. First, it is logically insufficient for the task. No single 2-D view can support accurate 3-D object recognition from multiple perspectives without further information (Vetter, Poggio et al. 1994). Second, if one assumes that other views are recognized by being rotated into correspondence with the canonical one, there must be some representation or process in memory to support rotation as previously occluded surface come into view. Third, the hypothesis of rotation from a single canonical view is inconsistent with the data from perspective experiments (Palmer and Bucher 1981; Edelman and Bülthoff 1992).

A more realistic possibility is that there might be multiple 2-D representations from several different viewpoints that can be employed in recognizing objects. The multiple-view representations need not actually be templates but could be feature maps or even structural descriptions, such as aspect graphs. An aspect graph (Koenderink and van Doorn 1979) is a network of representations containing all topologically distinct 2-D views (or aspects) of the same object. Each aspect is represented by a structural description that defines its topology: the set of edges and vertices that mark the borders between visible and occluded surfaces in the projected image. Aspect graphs improve on template theories of object representation by greatly reducing the number of view-specific representations required to capture the shape of object. They do so in part because, although they are view-specific representations, they are much more abstract than templates, being structural descriptions of the connectedness of edges and vertices and of the transitions between different views. Even so, the question remains of whether the reduction is great enough to make it a reasonable theory of human object recognition. A second difficulty with aspect graphs is that, because they are abstract, they cannot support discriminations among objects with different quantitative measurements that have the same 2-D topology of edges and vertices in their projected views. This is not a problem for geon representations because, despite largely qualitative nature, they include such as aspect ratio, symmetry and edge parallelism that are sensitive to metric qualities. A third problem with aspect graphs is that even if observers have seen only a single representative view of an object, they typically appear to have some idea of its 3-D structure and are able to recognize it from many novel views. This is an ability that is not directly supported by aspect graphs for a truly novel object, since they are essentially just networks of topological descriptions of the 2-D views from which that object has been seen. Although it is computationally possible to recognize novel views from a relatively small number of suitably constrained 2-D views (Ullman and Basri 1991).

In summary, all of structural description theories and view-specific theories have their own strengths and weaknesses. They are inadequate to capture the astonishing power, versatility, and subtlety of human object recognition. How people recognize objects is certainly among the most difficult problems in visual perception, so difficult that no satisfactory solution has yet been proposed. However, if a solution could be found to extract parts (i.e. geons or primitive volumetric components) from raw gray-scale images, therefore, to overcome the most difficult problem of structural description theories, the promising structural description approach may seem in explaining a wider range of phenomena in human visual perception than view-specific approaches do. In this dissertation, we developed a multi-layer neural network to automatically extract object parts from raw gray-scale images and finally achieve unsupervised object recognition within structural description framework.

However, when an object was rotated away from the training viewpoint, the part structural description of the object was altered accordingly. These changes may produce the generalization decrement of structural description models, thus causing problems for object recognition with rotational invariance. Integrating view-specific theory into structural description theory, we proposed a smart solution for this problem: training models with multiple views of an object to result in a more complete structural description, thereby allowing for recognizing object with rotational invariance.

1.3 Research Goal

The goal of this dissertation is to harvest the primate visual system to develop a series of visual models in hierarchical network architecture. The basic premise of this work is the hypothesis that the visual system achieves its robust performance by sparse coding of images in combination with unsupervised learning processes. Our goal is to demonstrate that a system based on these biologically motivated principles is able to achieve the desirable property of rotational invariance.

For a variety of primates, desirable objects and situations—such as nourishing food, protective shelter, and desirable mates—must be sought out and approached; dangerous objects and situations—such as precipitous drops, falling objects, and hungry or angry predators—must be avoided or fled from. Thus, to behave in an evolutionarily adaptive manner, we must somehow get information about what objects are present in the world around us, where they are located, and what opportunities they afford us. Given its biological importance, primate visual system must be evolved for millions of years to allow the survival and successful reproduction of organisms, therefore, provides a rich source of computational paradigms that can be used as inspiration for revolutionary visual models.

Biologically based principles such as sparse coding may have the information processing capabilities as well as huge payoffs in power/energy minimization and optimal resource management (Barlow 1994). Sparse coding is a distributed population coding in which very few cells are excited in a represented pattern. It has been shown that such an encoding strategy leads to superior information storage capacity compared to local (grandmother cell or Gnostic representations) or distributed information representation (Attwell and Laughlin 2001). Because only very few cells need to be activated and there are only a few cells encoding an event, sparse representations have intrinsic fault-tolerance and low-power consumption potential (Jabri 2000), which is a key requirement for any physical realization of computational models. Thus, sparse coding is important from an implementation perspective. Moreover, the physical connectivity of large scale networks requires strategies that exploit sparseness, local connectivity and population-based encoding of networks.

Developmental and adult plasticity are critical in primate vision. However, none of the information about the contents of scenes is available during learning (Edelman, Intrator et al. 2002). For instance there are around 10^6 photoreceptors in each eye whose activities are constantly changing with the visual world, which provide all the information that is available to indicate what objects there are in the world, how they are presented, what the lighting conditions are, etc. This makes unsupervised methods essential, and equally, allows them to be used as computational strategies for developing advanced visual modeling that is both highly sophisticated and versatile. Unsupervised learning studies how system can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input (Narendra and Parthasarathy 1990).

It took the computer vision community several decades to realize that attempting to reconstruct the visual world in the form of a detailed general-purpose 3D geometrical representation is a futile undertaking (Hinton and Sejnowski 1999). A more promising approach depends on the observation that images are not random but are generated by physical processes of limited complexity and that the appropriate response to an image nearly always depends on the physical causes of the image rather than the pixel intensities (Hinton, Dayan et al. 1995). This suggests that an unsupervised learning process should be used to solve the difficult problem of extracting the underlying causes, and decisions about responses can be left to a separate learning algorithm that takes the underlying causes rather than the raw sensory data as its inputs. Using the world as its own representation and extracting information from it, as necessary, through the action of feature detectors based on the notion of receptive fields(RFs), unsupervised learning allowed our visual models to automatically extract features of novel objects based on statistical properties of input images, and without employing any explicit knowledge.

Inspired by the primate visual ventral pathway, we propose in Chapter 2 a computational scheme for invariant object recognition within a structural description framework. Primates have a remarkable ability to interpret complex scenes in real time, despite the limited speed of the neuronal hardware available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, most likely to reduce the complexity of scene analysis. Consistent with the visual ventral pathway, we assume that the bottom-up saliency and top-town attention can complementarily filter out unwanted information from typically cluttered real-world scenes and to focus on what is important in a given situation, which would largely reduce the computational complexity of our visual models.

Chapter 3 develops V1 and V2 models for low-lever visual feature extraction with the sparse and unsupervised algorithms based on Independent Component Analysis (Bell and Sejnowski 1995). Taking the output of V1 layer as input, the RFs of the V2 model cells (M-cells) were trained by combining the sparse coding method with a non-negative constraint. The learned V2 RFs are selective for contour length in addition to being tuned to position and orientation, and exhibit end-stopping properties.

Compared with template matching, Fourier analysis and feature lists, the parts-based mechanism employed by our V4 model in Chapter 4 is better adapted to the real-world difficulties of view-point transforms, partial occlusion and plastic deformation. With the sparse coding constraint, this V4 model combines unsupervised representation in the feed-forward stream with lateral interaction to achieve stable, efficient and natural representation of shapes. Further comparing the response properties of V4 M-cells to those of biological cells (B-cells), we find that the M-cells display same curvature and object centered tuning as the reported physiological measurements.

During the past decade, there have been major advances in the understanding of how object recognition is performed in the primate visual system. There is now a large body of evidence showing that object recognition makes use of cells in inferior temporal cortex (IT) that respond to features of intermediate complexity. In Chapter 5, we present a method to train our IT model to do object detection from different viewing angles. We focus on a model where objects are represented as flexible constellations of rigid parts extracted by V4 model from raw gray-scale images. The variability within a class is represented by a joint probability density function (pdf) on the shape of the constellation. The set of model parameters, including the shape pdf, is then learned using expectation maximization (EM) algorithm. The method achieves very good classification results with approximate viewpoint invariance.

Chapter 2

A Biological Inspired Object Recognition System

2.1 Visual Information Processing in the Brain

Our visual sense provides us with rich and detailed information about the environment: we use it effortlessly to find out our way, recognize our friends and distinguish their moods, guide our movements, manipulate objects, and for a large variety of other tasks. Vision is a complex process that includes many interacting components, for example, with the analysis of color, depth, motion, shape, and texture of objects, and with the use of visual information for recognition, locomotion, and manipulation. The collection of processes involved in visual perception is often perceived as a hierarchy spanning the range from “low” via “intermediate” to “high-level” vision (Ullman 1996).

Low-level vision is usually associated with the extraction of certain physical properties of the visible environment, such as depth, edges, shape, object boundaries, or surface material properties. It is also considered “bottom-up” in nature. This means that they are determined by the data and are relatively independent of the task at hand or knowledge associated with specific objects. High-level vision, in contrast, is concerned with tasks such as the extraction of shape properties and spatial relations, and with object recognition and classification. It is concerned primarily with the interpretation and use of the information in the image, rather that the direct recovery of physical properties. High-level visual processes apply to a selected portion of the image, rather than uniformly across the entire image, and they often depend upon the goal of the computation and knowledge related to specific objects. Within the domain of higher-level vision, processes that do not depend on knowledge about specific objects are sometimes called intermediate-level vision. For example, the extraction of shape properties or object parts and spatial relations is often confined to selected elements in the scene and does not proceed uniformly and in parallel across the image. At the same time, the processing is largely independent of knowledge associated with specific objects.

It is also widely thought that any comprehensive processes of the visual information will be formed through the aggregation of components that each carries information about a particular aspect of the image. The kinds of component analyses most often discussed are those that alone might permit the grouping of parts of the image, as a precursor to the segmentation of surfaces in a scene. Among the dimensions suggested for independent analysis are luminance (brightness), texture, color, binocular disparity, movement, and orientation. The most influential modern evidence for independent analysis of visual attributes comes from physiological and anatomical studies of monkey visual cortex (Marr 1982; Livingstone and Hubel 1987; Lennie 1998).

2.1.1 The Visual Pathway

All visual information that the human mind receives is processed by a part of the brain known as visual cortex. The visual cortex (see Fig. 2.1) is part of the outermost layer of the brain, the cortex, and is located at the dorsal pole of the occipital lobe; more simply put, at the lower rear of the brain (Lennie 1998). The visual cortex obtains its information via projections that extend all the way through the brain from the eyeballs (Holmes 1918). The projections first pass through a stopover point in the middle of the brain, an almond-like lump known as the Lateral Geniculate Nucleus (LGN). From LGN the visual information are projected to the visual cortex for processing. Two different sorts of modular organization seem to be present in the visual cortex. One is represented by parallel pathways for the transmission of visual information (Phillips, Zeki et al. 1984); the other is represented by a hierarchy of stages of analysis (Wilson and Cragg 1967).

Visual cortex (Hubel 1963) is broken down into five areas, labeled V1, V2, V3, V4, and MT, which on occasion is referred to as V5. Each of these areas is further subdivided and sends information to any of 20 or more other areas of the brain that process visual information. This general arrangement is subdivided into two parallel pathways.

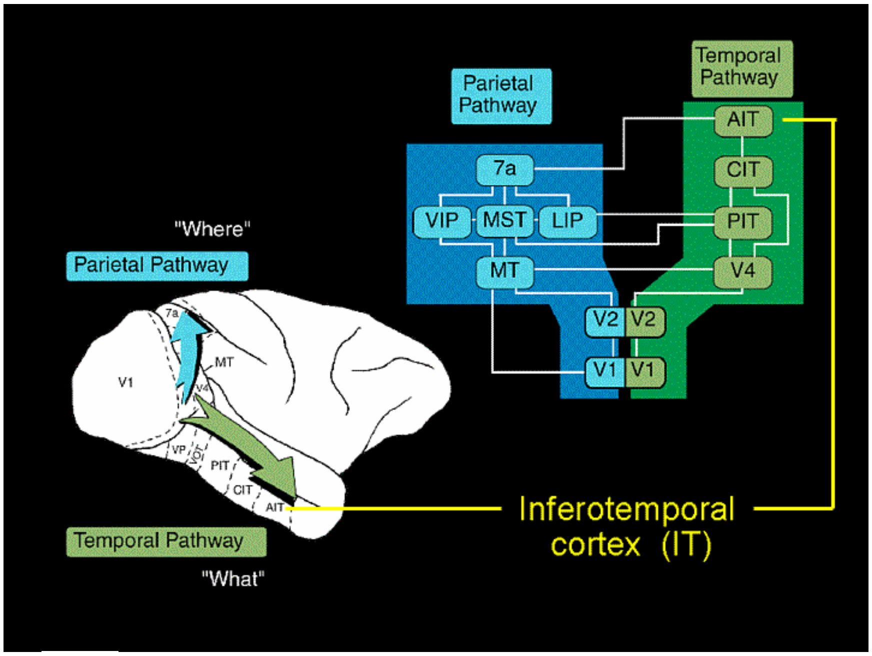


Figure 2.1 The visual cortex is located at the lower rear of the brain and broken down into five areas, labeled V1, V2, V3, V4, and MT (V5) (inspired by (Livingstone and Hubel 1987).

Although each pathway is somewhat distinct in function, there is intercommunication between them (Lennie 1998).

The ventral pathway (also called the “what” pathway) is: LGN parvo layers → V1 layer 4Cβ → V1 interblobs → V2 interstripes → V4 → IT, which is composed of feature detectors (simple, complex and hypercomplex cells). Parvocellular cells show a low sensitivity to contrast, high spatial resolution, and low temporal resolution or sustained responses to visual stimuli (Ungerleider and Mishkin 1982). These cellular characteristics make the parvocellular division of the visual system especially suited for the analysis of detail in the visual world. Cells found in the inferotemporal (IT) cortex, the highest purely visual area in the ventral visual stream and is thought to have a key role in object recognition, are tuned to views of complex objects such as faces: they discharge strongly to a face or face-like objects but very little or not at all to other objects. A hallmark of these cells is the robustness of their responses to stimulus transformations such as size/scale and position changes (Ito, Tamura et al. 1994). Damage to the parvocellualar pathway will induce disorders of object recognition

(Merigan 1996). Common examples of such disorders include visual agnosia, or the inability to identify objects in the visual realm, and prosopagnosia, a subtype of visual agnosia that affects specifically the recognition of once familiar faces. This division of the visual system tells us to identify **what** we see.

The dorsal pathway (also called "where" pathway) is: LGN mango layers→V1 layer 4Cα→V1 layer 4B→V2 thick stripes→V3→MT/V5, which an extension of the magnocellular pathway from the retina and LGN, and continues the processing of visual detail leading to the perception of shape in area V3 and movement or motion in areas V5 and MST (Zeki and Bartels 1998). Cells in V5 are particularly sensitive to small moving objects or the moving edge of large objects. Cells in dorsal MST respond to the movement (rotation) of large scenes such as is caused with head movements, whereas cells in ventral MST respond to the movement of small objects against their background (Graham 1989). Magnocellular cells show a high sensitivity to contrast, low spatial resolution, and high temporal resolution or fast transient responses to visual stimuli. These cellular characteristics make the magnocellular division of the visual system especially able to quickly detect novel or moving stimuli, the abilities that allow us to respond quickly and adaptively to possible threatening stimuli.

It should be kept in mind that within these ventral and dorsal streams of visual processing, visual cells are not passively responding to visual stimulation. Their responses are constantly modified by selective attention (Treisman 1986).

2.1.2 The Visual Receptive Fields

The receptive field (RF) of a sensory cell is a region of space in which the presence of a stmulus will alter the firing of that cell. RFs have been identified for cells of the visual system, the auditory system, and the somatosensory system .

RFs are probably the most ubiquitous computational mechanism employed in biological information processing (Boussaoud, Desimone et al. 1991). In the visual system, the RF of a cell is defined as the part of the visual field in which a stimulus must appear to elicit a response from the cell. The layered nature of the cortical architecture and the peculiarities of the inter-layer connections (Hubel 1995) that ultimately give rise to the profiles of individual RFs together constrain the kind of information processing that can be supported by the cortex.

The notion of an RF was put forward when the first recording of cell activity in primitive visual systems showed that cells only respond when the stimulus (usually a small spot of light) is confined to a well-defined region of the visual field (Hartline 1938). Subsequently, it was found that not all portions of an RF contribute equally to the generation of a response, leading to the definition of an RF profile. For example, the RF of a retinal ganglion cell in vertebrates is composed of a central region, surrounded by an annulus whose contribution to the response of the entire cell is opposite in sign to that of the center.

Hubel and Wiesel (1963) classified RFs of cells in the V1 cortex into simple cells, complex cells, and hypercomplex cells. Simple cell RFs are elongated, for example, with an excitatory central oval, and an inhibitory surrounding region, or approximately rectangular, with one long side being excitatory and the other being inhibitory. Images for these RFs need to have a particular orientation in order to excite the cell. For complex-cell RFs, a correctly oriented bar of light might need to move in a particular direction in order to excite the cell. For hypercomplex RFs, the bar might also need to be of a particular length.

As one moves through the ventral stream that underlies object recognition, the receptive field size of cells increases steadily (see Fig. 2.2). At corresponding eccentricities near the fovea receptive fields in V2 are (in linear dimensions) 2-3 times larger than in V1; in V4 perhaps 5-6 times larger; cells in IT have receptive fields that can include the entire central visual field, on both sides of the vertical midline (Wallis and Rolls 1997). These large receptive fields are presumably necessary to recognize large complex objects and may mediate the ability to recognize objects of any size as the same, regardless of their retinal location (Smith, Singh1 et al. 2001).

RFs constitute an important variation on the theme of computing with connections, which is central to many distributed models of perceptual and cognitive functions. In

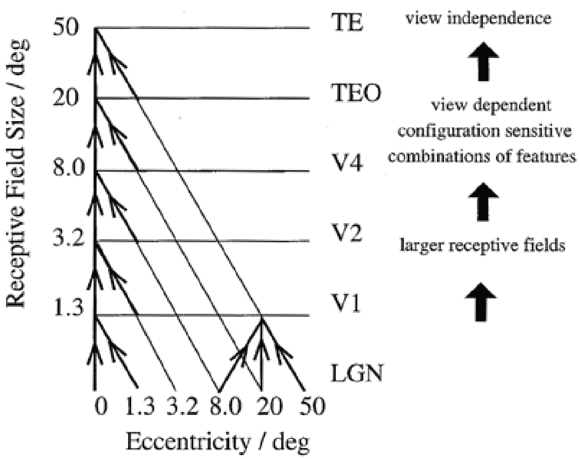


Figure 2.2 Schematic diagram showing convergence achieved by the forward projection in the ventral pathway(Wallis and Rolls 1997). Based on the size of the receptive fields and the properties of response, the inferotemporal cortex (IT) is divided into TEO and TE areas.

sensory physiology, the notion of functional units selectively and reliably responsive to external stimuli (that is, units whose output co-varies with the sensory input in a well-defined and predictable manner) has found a wide acceptance in the form of the feature detector doctrine, which is intimately linked to the concept of RF. Confining the input to a unit to a localized compact subset of units in the preceding layer addresses the problem of natural encoding of the topology of images, but does not solve the problem of the complexity of learning in networks with too many parameters. A solution to this problem calls for making the weight pattern within each RF the same, resulting in a drastic reduction in the number of independent parameters in the model (LeCun and Bengio 1995). The action of the network on the image becomes then equivalent to a convolution with a bank of translationally invariant RFs. The convolution network resulting from this so-called weight sharing thus occupies an intermediate position between two extremes: hand-tailed connection patterns on the one hand, and universal function approximation schemes on the other hand. In a word, the effectiveness of living perceptual systems (Winger 1960) stems from a peculiar match between statistical and parametric characterizations of visual objects, and properties of visual RFs.

2.1.3 Bottom-up and Top-down Bias

Primates have a remarkable ability to interpret complex scenes in real time, despite the limited speed of the neuronal hardware available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, most likely to reduce the complexity of scene analysis. This selection appears to be implemented in the form of a spatially circumscribed region of the visual field, the so-called “focus of attention”, which scans the scene both in a rapid, bottom-up, saliency-driven and task-independent manner as well as in a slower, top-down, volition-controlled and task-independent manner (Itti, Koch et al. 1998).

Bottom-up Saliency: Saliency, eclectically defined by visual attributes such as color, intensity, orientation, direction of movement, as well as the contrasts of those attributes, is used by the biological vision system to provide a first-order approximation of the area of interest in visual scene when it is either viewed without any prior knowledge or is not given any attention (Gottlieb, Kusunoki et al. 1998). Understanding this mechanism is important because bias is likely to be deployed, during the first few hundred milliseconds after a new scene is freely viewed, mainly based on bottom-up cues.

Some stimuli are intrinsically conspicuous or salient in a given context. For example, a red dinner jacket among black tuxedos at a somber state affair, or a flickering light in an otherwise static scene, automatically and involuntarily attract attention. Saliency is independent of the nature of the particular task, operates very rapidly and is primarily driven in a bottom-up manner, although it can be influenced by contextual, figure-ground effects. If a stimulus is sufficiently salient, it will pop out of a visual scene. This suggests that saliency is computed in a pre-attentive manner across the entire visual field. The speed of this saliency-based form of attention is on the order of 25 to 50ms per item.

The bottom-up saliency biases the competition of cells in the ventral stream. So that, primates can detect, in real time, conspicuous objects in cluttered visual environments. Reproducing such non-specific target detection capability in artificial systems has important applications, for example, in embedded navigational aids, in robot navigation and in battlefield management. Based on psychophysical studies in humans and electrophysiological studies in monkeys, it is believed that bottom-up visual saliency acts in some way akin to a “spotlight” (Gottlieb, Kusunoki et al. 1998). The spotlight can rapidly shift across the entire visual field (with latencies on the order of 50ms), and select a small area from the entire visual scene. The neuronal representation of the visual world is enhanced within the restricted area of the attentional spotlight, and only this enhanced representation is allowed to progress through the cortical hierarchy for high-level processing, such as pattern recognition. Further, psycholophysical studies suggest that only this spatially-circumscribed enhanced representation reaches visual awareness and consciousness.

Low-level feature extraction mechanisms act in a massively parallel manner over the entire visual scene to provide bottom-up biasing cues towards salient image locations. The salient image locations then sequentially pop-out to be analyzed in more detail. Visual saliency hence allows for seemingly real-time performance by breaking down the complexity of scene understanding into a fast temporal sequence of localized pattern recognition problems.

Several models have been proposed to functionally account for many properties of visual saliency and/or attention in primates (Olshausen, Anderson et al. 1993; Itti, Koch et al. 1998; Niebur and Koch 1998; Itti and Koch 2000; Itti and Koch 2001). These models typically share similar general architecture. Multi-scale topographic “feature maps” detect local spatial discontinuities in intensity, color, orientation, and optical flow. In biologically-plausible models, this is usually achieved by using a “center-surround” mechanism akin to biological visual RFs, a process also known as a “cortex transform” in the image processing literature. RF properties can be well approximated by difference-of-Gaussians (DOG) filters (for non-oriented features) or Gabor filters (for oriented feature)(Itti, Koch et al. 1998). Feature maps from different visual modalities are then combined into a unique “master” or “saliency” map. In models like, presumable, in primates, the saliency map is sequentially scanned, in order of saliency, by the focus of attention (see Fig. 2.3). A central problem, both in biological and artificial systems, is that of combining multi-scale feature maps, from different visual modalities with unrelated dynamic ranges (such as color and motion), into a unique saliency map. In research done by Itti (Itti and Koch 2000), models usually assume simple summation of all feature maps, or a linear combination using ad-hoc weights. Overall, simple normalized summation always yielded poorest performance. Linear combination with weights learned by supervised additive training yielded the best performance with a fast time to find a salient target, but with poor generalization. A non-linear combination which enhances feature maps with a few isolated peaks of activity, while suppressing feature maps with uniform activity, yielded significantly better performance with improvement in time, while preserving generality. The strategy of local, non-linear competition between salient locations followed by summation yielded the better performance too (Itti and Koch 2001).

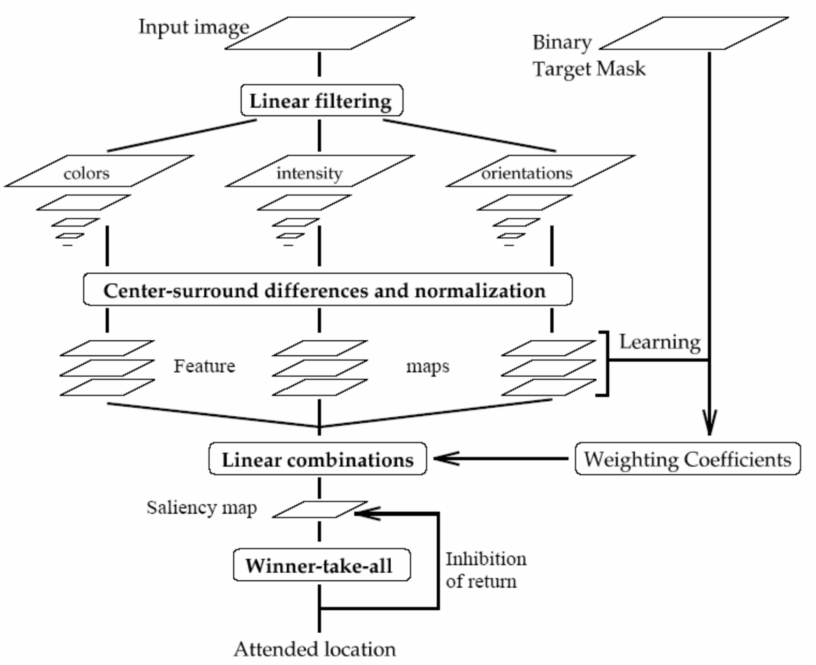


Figure 2.3 The general architecture of the visual saliency system. Early visual features are extracted in parallel in several multiscale "feature maps", which represent the entire visual scene. Such feature extraction is achieved through linear filtering for a given feature type (e.g. intensity, color or orientation), followed by a center-surround operation which extracts local spatial discontinuities for each feature type. All feature maps are then combined into a unique "saliency map". After such combination is computed, a maximum detector selects the most salient location in the saliency map and shifts attention towards it. This location is subsequently suppressed (inhibited), to allow the system, to focus on the next most salient location (Itti and Koch 2001).
Figure

Top-down attention: Even while our eyes are fixated on a particular location, it does not appear that the visual system passively processes all the information available within the image. Rather, we selectively attend to different aspects of it at different times. Sometimes we attend globally to the whole scene; at other times we attend to a selected object or set of objects; at still other times we attend locally to a specific object part. We may even concentrate on a particular object, such as the color or texture of a shirt we are considering buying. Our ability to engage in these flexible strategies for processing different information within the visual field-generally referred to as attention – is therefore an important component of vision. Overt eye movements determine what optical information is available to the visual system; Covert selective attention determines what subset of this information gets full processing.

Top-down attention is a more deliberate and powerful one that has variable selection criteria, depending on the task at hand (for example, look for the red, horizontal target'). The expression of this top-down attention is most probably controlled from higher areas, including the prefrontal cortex (PFC), which connects back into visual cortex and early visual areas (Luck, Chelazzi et al. 1997; Reynolds, Chelazzi et al. 1999; Kanwisher and Wojciulik 2000; Reynolds and Desimone 2003). Such volitional deployment of attention has a price, because the amount of time that it takes – 200ms or more – rivals that needed to move the eyes. So, whereas certain feature in the visual world automatically attract attention and are experienced as "visually salient", directing attention to other locations or objects requires voluntary "effort". Both mechanisms can operate in parallel.

The top-down attention poses a mechanism for the selection of behaviorally relevant information from natural scenes which usually contain multiple objects. The limited processing capacity of the visual system does not allow the simultaneous analysis of many different objects. Selective visual attention facilitates therefore the processing of that limited portion of the input associated with the relevant information and it suppresses the remaining irrelevant information. If attention is, for example, top-down guided to a certain spatial location in the visual field, as in the case of object recognition, information coming from the attended location will be facilitated whereas the information coming from the other unattended locations will be suppressed. On the other hand, if the top-down bias the attention in such a way that only a specific feature (e.g. color) or a bundle of conjoined features (e.g. a given object) are important, as in the case of visual search, then only the processing of the corresponding consistent input features in the visual field is augmented whereas the remaining irrelevant feature information is suppressed.

Consistent with the existing experimental single cell recordings (Reynolds and Desimone 1999), Deco proposed a multimodular neurodynamical system for top-down attention based on the “biased competition” hypothesis (Duncan 1980). The “biased competition” hypothesis proposes that multiple stimuli in the visual field activate populations of cells that engage in competitive interactions. On the other hand, attending to a stimulus at a particular location or with a particular feature biases this competition in favor of cells that respond to the features or location of the attended stimulus. This attentional effect is produced by generating signals within areas outside visual cortex which are then fed back to extrastriate areas, where they bias the competition such that when multiple stimuli appear in the visual field, the cells representing the attended stimulus "win", thereby suppressing cells representing distracting stimuli.

Deco’s system (Deco 2001) is essentially composed of three modules structured such that they resemble the two known main visual pathways of the primate visual cortex (see Fig. 2.4). Information from the retino-geniculo-striate pathway enters the visual cortex through area V1 in the occipital lobe and proceeds into two processing streams. The occipital-temporal stream (the so-called "what" pathway) leads ventrally through V2, V4 and IT and is mainly concerned with object recognition, independently of position and scaling. The occipito-parietal stream (the so-called "where" pathway) leads dorsally into posterior parietal (PP) complex and is concerned with the location and spatial relationships between objects. In the system, object recognition and visual search can be explained in the theoretical framework of a biased competitive neurodynamics. The top-down bias can guide attention to concentrate at a given spatial location or at given features. The neural population dynamics are handled in the framework of the mean-field approximation, i.e. by the analytical representation of the mean activity of a population of cells. Consequently, the whole process can be expressed as a system of coupled differential equations.

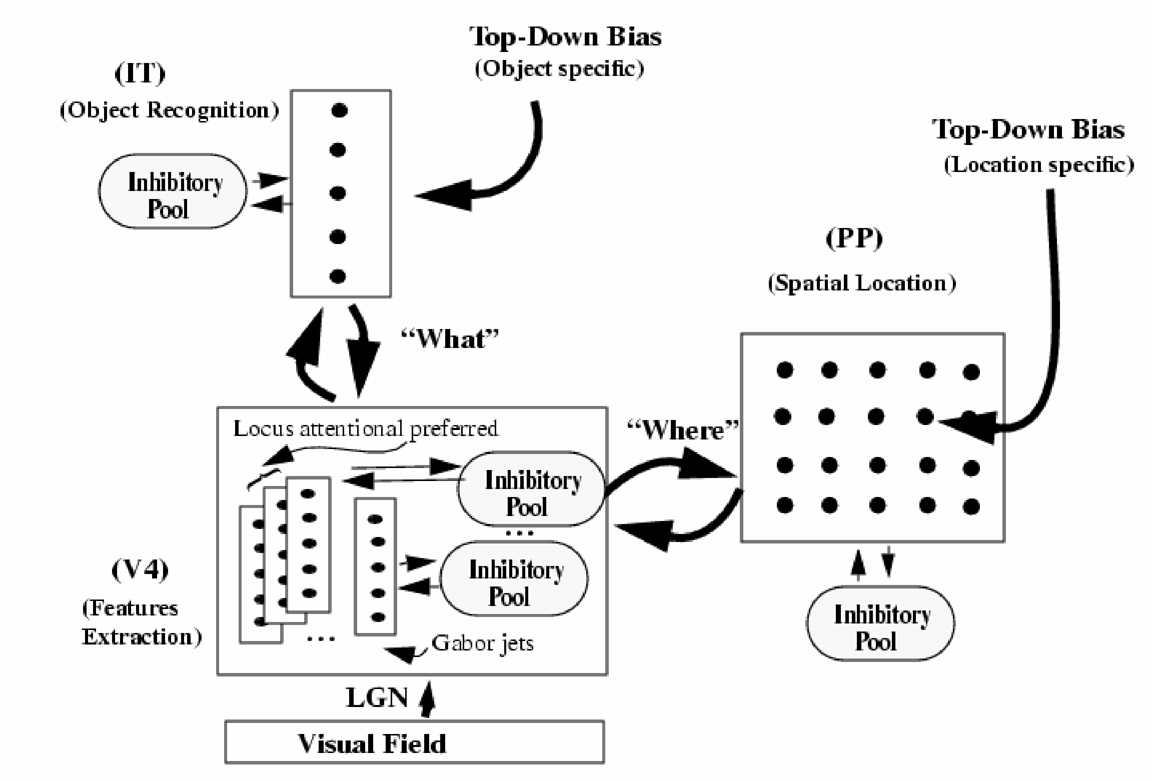


Figure 2.4 A multimodular neurodynamical system for top-down attention based on the "biased competition" hypothesis (Deco 2001).

2.2 Unsupervised Learning

Unsupervised learning techniques deal with how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input (Barlow 1989); rather, the unsupervised learner brings to bear prior biases as to what characteristics of the structure of the input should be captured in the output. In other words, there are no specific examples of the function to be learned by the network. Instead, provision is made for a task-independent measure of the quality of the representation that the network attempts to learn. The free parameters of the network are then optimized with respect to that measure. Once the network has become tuned to the statistical regularities of the input data, it develops the ability to form internal representations for encoding features of the input and thereby create new classes automatically (Becker 1991).

2.2.1 Methods for Unsupervised Learning

Two classes of method have been suggested for unsupervised learning. The first, density estimation techniques explicitly build statistical models (such as Bayesian networks) of how underlying causes could create the input, i.e. build generative models that are likely to have produced the observed data. The parameters of these generative models are adjusted to optimize the likelihood of the data within constrains such as basic assumptions about the model architecture. The second, feature extraction techniques try to extract statistical regularities (or sometimes irregularities) directly from the inputs.

Unsupervised learning has a long and distinguished history. Some early influences included the work of Horace Barlow (Barlow 1989), who sought ways of characterizing neural codes, Donald MacKay (MacKay 1956), who adopted a cybernetic-theoretic approach, and David Marr (Marr and Nishihara 1978), who made an early unsupervised learning postulate about the goal of learning in his model of the neocortex. The Hebb rule (Hebb 1949), which links statistical methods to neurophysiological experiments on plasticity, has also cast a long shadow. Geoffrey Hinton and Terrence Sejnowski (Hinton and Sejnowski 1986) inventing a model of learning called the Boltzmann machine, imported many of the concepts from statistics that now dominate the density estimation methods.

Maximum likelihood density estimation, and approximations to it, covers a very wide spectrum of the principles that have been suggested for unsupervised learning. This includes versions of the notion that the outputs should convey most of the information in the input; that they should be able to reconstruct the inputs, perhaps subject to constraints such as being independent or sparse; and that they should report on the underlying causes of the input. Many different mechanisms apart from clustering have been suggested for each of these, including various forms of Hebbian learning (Hebb 1949), the Boltzmann and Helmholtz machines (Sejnowski and Rosenberg 1987), sparse-coding, various other mixture models, and independent components analysis (Bell and Sejnowski 1995).

Density estimation is a heuristic for learning good representations. It can be too stringent — making it necessary to build a model of all the irrelevant richness in sensory input. It can also be too lax — a look-up table that reported $P_I(X)$ for each X might be an excellent way of modeling the distribution, but provides no way to represent particular examples of X .

Another smaller class of unsupervised learning methods seeks to discover how to represent the inputs X by defining some quality that good features have, and then searching for those features in the inputs. For instance, consider the case that the output $Y = W \cdot X$ is a linear projection of the input onto a weight vector . The central limit theorem implies that most such linear projections will have Gaussian statistics. Therefore if one can find weights such that the projection has a highly non-Gaussian (for instance, multi-modal) distribution, then the output is likely to reflect some interesting aspect of the input. This is the intuition behind a statistical method called projection pursuit. It has been shown that projection pursuit can be implemented using a modified form of Hebbian learning (Intrator 1992). An arrangement that different outputs represent different aspects of the input turns out to be surprisingly difficult to do.

Projection pursuit can also execute a form of clustering. Consider projecting the photoreceptor activities onto the line joining the centers of the clusters. The distribution of all activities will be bimodal – one mode for each cluster – and therefore highly non-Gaussian, though this single projection does not characterize well the nature or shape of the clusters.

Another example of heuristic underlying good features is that causes are often somewhat global. For instance, consider the visual input from an object observed in depth. Different parts of the object may share a few features, except that they are at the same depth, *i.e.* one aspect of the disparity in the information from the two eyes at the separate locations is similar. This is the global underlying feature. By maximizing the mutual information between outputs y_i and y_j that are calculated on the basis of the separate input, one can find this disparity. This technique was invented by Becker & Hinton (1992) and is called IMAX.

2.2.2 Generative Models

The large class of unsupervised learning methods consists of maximum likelihood (ML) density estimation methods. All of these are based on building parameterized models $P[X;\Phi]$(with parameters Φ) of the probability distribution $P_I(X)$ where the forms of the models (and possibly prior distributions over the parameters) are constrained by a priori information in the form of the representational goals. These are called synthetic or generative models, since, given a particular value of Φ, they specify how to synthesize or generate samples from $P[X;\Phi]$, whose statistics should match $P_I(X)$. A typical generative model has the structure:

$$P[X;\Phi] = \sum_{Y} P[Y;\Phi]P[X \mid Y;\Phi] \tag{2.1}$$

where Y represents all the potential causes of the input X. The typical measure of the degree of mismatch between $P[X;\Phi]$ and $P_I(X)$ is called the Kullback-Leibler divergence:

$$KL[P_I[X],P[X;\Phi]] = \sum_{X} P_I[X]\log\left[\frac{P_I[X]}{P[X;\Phi]}\right] \geq 0 \tag{2.2}$$

with equality if and only if $P_I[X] = P[X;\Phi]$.

Given an input pattern , the most general output of this model is the posterior, analytical, or recognition distribution $P[Y/X;\Phi]$, which recognizes which particular causes might underlie X. This analytical distribution is the statistical inverse of the synthetic distribution.

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

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

2.3 Sparse Coding

At any given moment, our senses are receiving vast amounts of information about the environment in the form of light intensities, changes in sound pressure, deformations of the skin, stimulation of taste and olfactory receptors and more. How the brain makes sense of this flood of time-varying information and forms useful internal representations for mediating behavior remain one of the outstanding mysteries in neuroscience. In recent years, a combination of experimental, computational, and theoretical studies (Amari 1993; Barlow 1994; Field 1994; Simoncelli 2003) have pointed to the existence of a common underlying principle involved in sensory information processing, namely that information is represented by a relatively small number of simultaneously active cells out of a large population, commonly referred to as 'sparse coding'.

Horace Barlow (Barlow 1989) has argued that statistical redundancy contains information about the patterns and regularities of sensory stimuli. Completely non-redundant stimuli are indistinguishable from random noise, and Barlow claims that the percept of structure is driven by the dependencies. According to Barlow's theory, what is important is that a system detects new statistical regularities in the sensory input that differ from the environment to which the system has adapted. Barlow termed these new dependencies "suspicious coincidences." Bars and edges, for example, are locations in the visual input where there is phase alignment across multiple spatial scales, and therefore constitute a “suspicious coincidence”.

Two strategies (see Fig. 2.5) take advantage of the redundancy in the environment to produce more effective representations. The first is described as a “compact” coding scheme. A compact code performs a transform that allows the input to be represented with a relatively small number of vectors (e.g., cells) with minimal loss in the representation of the input. In such a code, e.g. Principal Component Analysis (PCA), the dimensionality of the representation is reduced, resulting in a code where only a subset of the possible inputs can be accurately represented. The code is effective when this subset is capable of representing the probable inputs to the code.

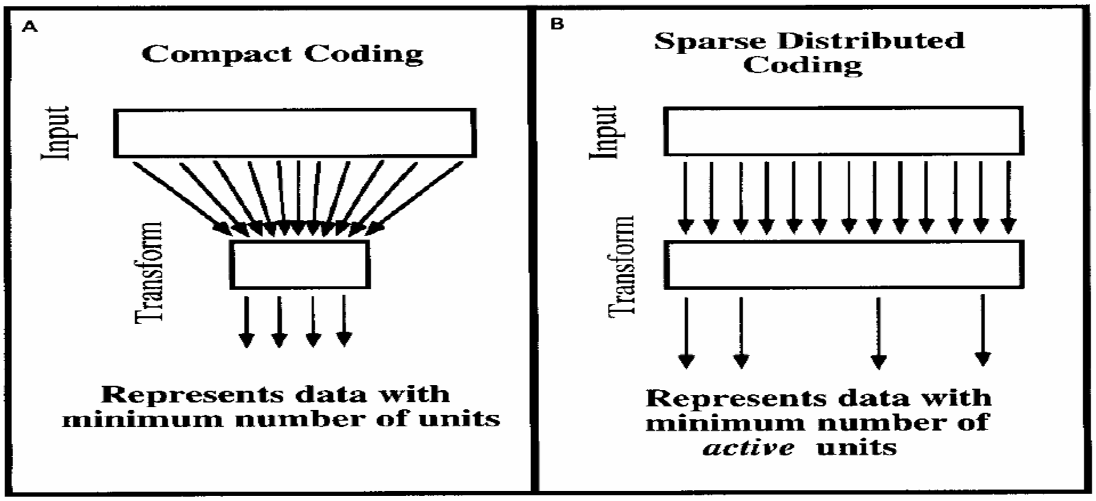

A
Compact Coding
Input
Transform
Represents data with minimum number of units
B
Sparse Distributed Coding
Input
Transform
Represents data with minimum number of *active* units

Figure 2.5 Two methods of taking advantage of redundancy in a sensory environment. In a compact coding scheme (A), the code transforms the vector space to allow the data to be repented by a smaller number of vectors (i.e. dimensionality is reduced) with only minimal loss in the representation. In a sparse coding scheme (B), the code transforms the vector space to allow the input to be represented with a minimum number of active cells. In a sparse coding scheme, the dimensionality is not reduced. Rather, the redundancy in the input is transformed into the redundancy in the firing rate of the cells (i.e., the response histogram) to produce a code where the response probability for any particular cell is relatively low (Field 1994).

The second scheme suggests that the principal goal of sensory coding is to produce a sparse-distributed representation of the sensory input. In a sparse-distributed code, the dimensionality of the representation is maintained (the number of cells remains roughly constant and may even increase). However, the number of cells responding to any particular instance of the input is minimized. Over the population of likely inputs, every cell has the same probability of producing a response (i.e., distributed) but the probability is low for any given cell (i.e. each representation is sparse). In such a code, the dimensionality is not reduced. Rather, the redundancy of the input is transformed into the redundancy of the firing pattern cells. The goal of the coding is to maximize the redundancy of the response histograms by minimizing the statistical dependencies between units.

In particular, an effective sparse code must have two properties. It must span the space of inputs (i.e. preserve information) and show high kurtosis in response histograms (i.e. a sparse representation is consistent with a response with a distribution that is highly peaked at zero and with a heavy tails compared to a Gaussian of the same variance). The kurtosis measures the 4th moment relative to the variance squared:

$$k = \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i - \overline{r})^4}{\delta^4} - 3 \tag{2.3}$$

where r is the response of the cell, $\overline{r}$ is the mean response and δ is its standard deviation. For a Gaussian (non-sparse) distribution $k = 0$, whereas for a heavy-tailed (sparse) distribution $k > 0$. In addition, measures such as kurtosis are extremely sensitive to outliers—that is, a small number of data points far from the mean can have a significant impact on kurtosis (Olshausen and Field 2004).

A compact code allows the data to be stored and transmitted with a smaller population of total cells. But what are the advantages of sparse codes?

First, a sparse coding scheme can increase the signal-to-noise ratio. If most of the variance of a data set is represented by a small subset of cells, then that subset must have a high response relative to the cells that are not part of the subset. The smaller the subset, the higher the response of each member of the subset, given an image with constant variance. If we consider the response of the subset as “the signal” and if all the cells in the population are subject to additive noise, then by increasing the response of the subset of cells relative to the population, it is possible to increase the probability of detecting the correct subset of cells that represent the image.

Second, although signal-to-noise ratio considerations may be important, it is proposed that the main reason for sparse coding is that it assists in the process of recognition and generalization. In an ideal sparse code, the activity of any particular basis function has a low probability. Because the response of each cell is relatively rare, tasks that matching of features should be more successful. As a code becomes more sparse, the probability of detecting a correct correspondence increases. This suggests that a sparse code should be helpful in tasks requiring solutions to a correspondence problem and can be related to Barlow's suspicious coincidence (Barlow 1989). If the probability that any cell responding is low ($p \leq 0.5$), then the probability of two cells responding is even lower as p^2 ($p \leq 0.25$) assuming response independence. Higher-order relations (pairwise, triad, etc.) become increasingly rare and therefore more informative when they are present in the stimulus.

This implies that the unique pattern of activity found in sparse codes may also involve in solving the general problem of recognition. In relations among units are used to recognize the particular view of an object, then with a sparse code, any particular higher-order relations is relatively rare. In a compact code, a few cells have relatively high probability of response. Therefore, any particular higher-order relation among this group is relatively common. In a compact code, different objects are represented in terms of the differential firing of the same subset of cells. With a sparse-distributed code, a large population has a relative low probability of response. Different objects are represented by unique subsets of cells. That is, different objects are represented by which cells are active rather than how much they are active.

The third advantage of sparse coding comes from research on networks with associative memories. Several authors have noted that when the inputs to the networks are sparse, the networks can store more memories and provide more effective retrieval with partial information (Palm 1980; Baum, Moody et al. 1988). Baum and his colleagues suggest that the advantages of sparsity for cell efficiency are "so great that it can be useful to artificially 'sparsify' data or responses which are not sparse to begin with". Indeed, it is not surprising that many types of networks will solve problems more efficiently if the inputs are first "sparsified". Since the sparse representation will have fewer higher-order relations, learning to classify or discriminate inputs should require less computation. Therefore, “sparsifying” the input should help to simplify many of the problems that the network is designed to face.

The fourth reason for favoring sparse codes is that they are energy efficient (Levy and Baxter 1996). Attwell and Laughlin (Attwell and Laughlin 2001) have recently produced an estimate of the energy required for signaling in cortical cells, and they conclude that average firing rates must be rather low, that is, less than 1 Hz. On the basis of their findings, Lennie (Lennie 2003) estimates that at any given moment only 1/50th of the population of cortical cells could afford to be significantly active. Thus, given the actual energy constraints of the mammalian cortex, sparse coding would seem to be a necessity.

2.4 Our Proposed Scheme for Invariant Object Recognition

Objects are usually not seen in isolation, but embedded in larger scenes, next to, and sometimes partially occluded by, other objects. The task becomes much more complicated when the object is not isolated, but “buried” within the context of a larger scene. To recognize an individual object in the scene, it would be useful to somehow separate as much as possible the image of the object in question from the remainder of the scene. A common approach has therefore been to precede the recognition process by a stage called “segmentation”. The goal of the segmentation stage is to isolate from the image a sub-structure that corresponds to the image of a single object. If such a goal can be realized, then the task of the subsequent recognition process will be simplified.

Simulated with the primate brain, we proposed a scheme (see Fig. 2.6) to break down the problem of understanding a visual scene into a rapid series of computationally less demanding, localized visual analysis problems; therefore, to facilitate visual information processing in the ventral pathway; finally, to achieve invariant object recognition by unsupervised and sparse learning.

The scheme is essentially composed of three modules: the visual ventral pathway, the bottom-up saliency and the top-down attention. Using bottom-up, image-based saliency cues (Itti’s model) and top-down, task-dependent cues (Deco’s model and Hamker’s model), the visual information processed in the ventral pathway focuses on what is important in a given situation, which will largely reduce the computational complexity and save the system memories.

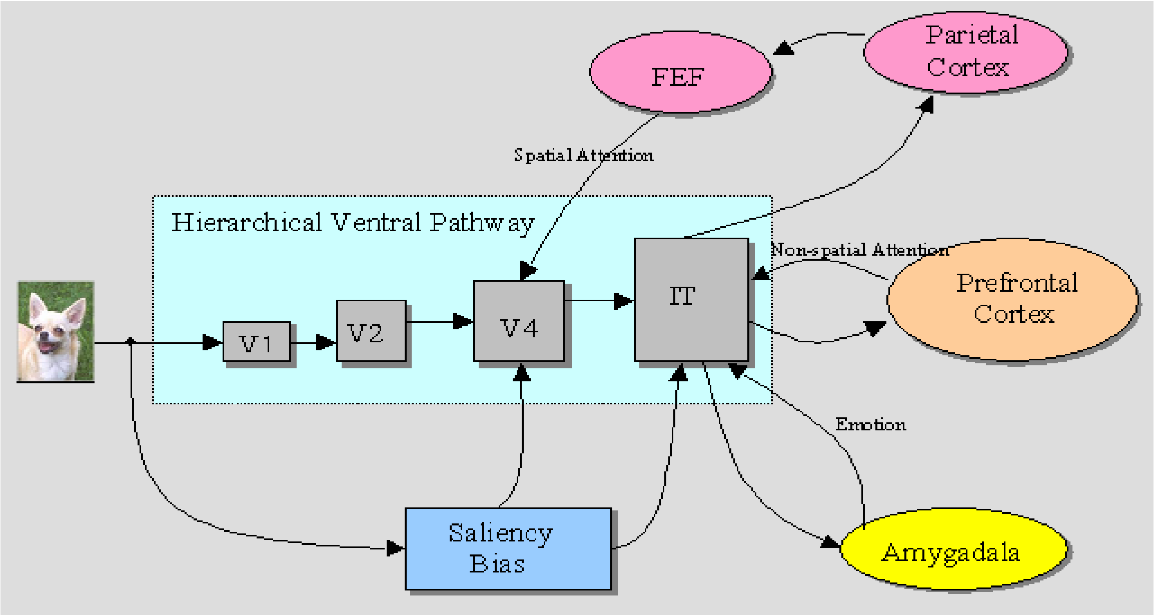


Figure 2.6 The proposed scheme uses bottom-up saliency and top-down attention to filter out unwanted information from typically cluttered real-world scenes, which facilitates visual information processing in the ventral pathway.

When we look at a scene without any prior knowledge or task, our attention will be attracted to some locations mostly because of their saliency, defined by contrasts in color, intensity, orientation etc. We postulate that saliency-based bias can roughly segment the objects of interest from the complex scene. The model extracts feature maps for orientation, intensities and color, and builds up a saliency map using intermediary conspicuity maps. All locations within the saliency map compete in a winner-take-all (WTA) network of integrate and fire cells, and the winning location is attended to. Subsequently, the saliency map is inhibited at the winning location (inhibition-of-return), allowing the competition to go on, so that other locations can be attended to. The goal of saliency bias is to provide the recognition system with a first order approximation of the location and extent of interesting objects in the scene. For this, the algorithm first determines which of the feature domains (colors, intensities, or orientations) contributed most to the saliency of the current focus of attention (FOA). Then the feature map that contributed most to the winning conspicuity map at the FOA is found. The winning feature map is segmented using a flooding algorithm with adaptive thresholding. The segmented feature map is used as a template for object-based inhibition-of-return of the saliency map. It is also processed into a binary mask and convolved with a Gaussian kernel to yield the modulation mask at image resolution. Finally, the modulation mask is multiplied to the activities of the V4 cells in the recognition system.

After information about the content and its low level stimulus-driven salience is extracted. This information is sent further upwards to V4 and to IT cells which are broadly tuned to location. The target template is encoded in working memories in prefrontal cortex (PF). Then, cells in PF are indicated by comparison of the target [1]template with the output of IT: whether the target is actively encoded in IT. Feedback from PF to IT increases the strength of all features in IT matching the template. Feedback from IT to V4 sends the information about the target downwards to cells with a higher spatial tuning. Frontal eye field (FEF) combines the feature information across all dimensions and indicates salient or relevant locations in the scene. A winner-take-all process in FEF cells selects the strongest location. Even during this competition a reentry signal from this map to V4 and IT enhances all features at locations of activity in FEF. The inhibition-of-return map memorizes recently visited locations and inhibits the FEF cells (Hamker and Worcester 2002).

Based on the assumption[2] that the competition among cells in the ventral pathway is biased by the bottom-up saliency and top-down attention, which means interesting object areas have been segmented from a complex visual scene, our research work focus on developing the low-level feature extraction (V1 and V2) and high-level shape representation (V4) and object recognition IT) models.

[2] Because the previous works by Itti and his colleagues can be easily integrated into our “ventral pathway”-like object recognition model, we will not repeat their works in our research

Chapter 3

Low-level Visual Feature Extraction

3.1 Introduction

3.1.1 V1 Cells

V1, sometimes called striate cortex because of its stripy appearance when dyed and put under a microscope, is by far the largest and most important. It is sometimes called primary visual cortex or area 17. The other visual areas are referred to as extrastriate cortex (Hubel 1963).

V1 is organized in columns (see Fig. 3.1). The term "cortical column" refers to the notion that cells arranged vertically from the surface of the cortex to the white matter might comprise functional or anatomical units. Thus, a cortical column can be defined on the basis of anatomical features (e.g. stereotyped patterns of pyramidal cell

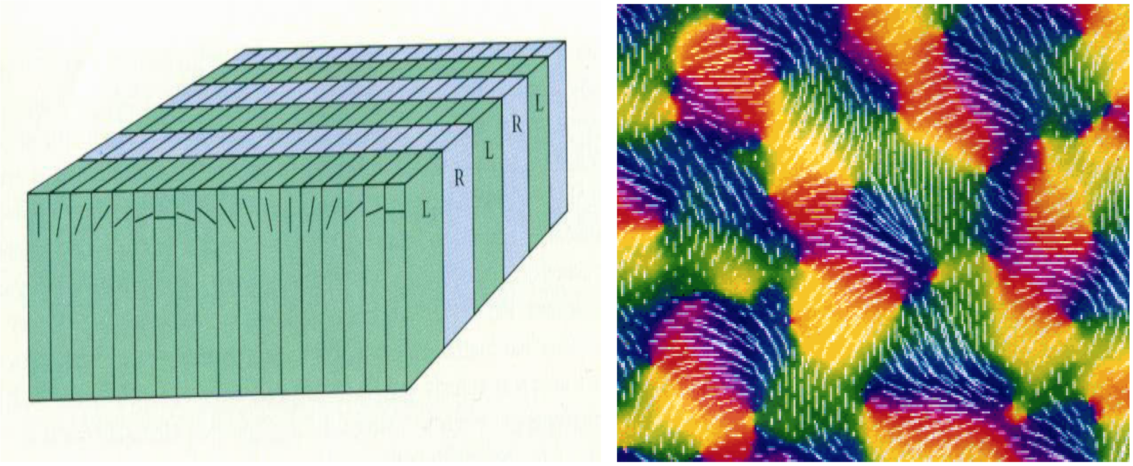


Figure 3.1 Orientation Columns of V1. A striking aspect of V1 is columnar organization of preferred stimulus orientation – all cortical cells in each column have similar preferred orientation; the columnar organization seen in V1 is believed to optimize encoding and visual analysis (Hubel 1995).

apical dendrite bundles), functional features (e.g. columns of cortical cells all responding to the same stimulus orientation) or both. All the cells in a given column have overlapping and similar receptive field (RF) properties. Adjacent columns in the visual cortex have adjacent RFs and differ in a systematic way in their RF properties.

Cells in V1 have elongated RFs, and consequently respond best to elongated stimuli, namely bars and edges.

Simple cells in the primary visual cortex have RFs which are restricted to small regions of space and are highly structured. The size of a simple-cell RF depends on its position in the retina relative to the fovea. The smallest fields, in and near the fovea, are about one-quarter degree by one-quarter degree in total size of a visual field. In the far retinal periphery, simple-cell RF can be about 1 degree by 1 degree.

Simple cells have clearly defined excitatory and inhibitory regions. Simple cell RFs contain sub-regions that exert excitatory and inhibitory influence on the cell's response; The most effective stimulus is one that puts most light in the excitatory region, and only little in the inhibitory region. In addition, the light pattern must have the right orientation, position and size. Simple cell RFs can be 'built' in the cortex by collecting responses from LGN cells whose RFs fall along a line across the retina, but the exact wiring is still the subject of debate (see Fig. 3.2).

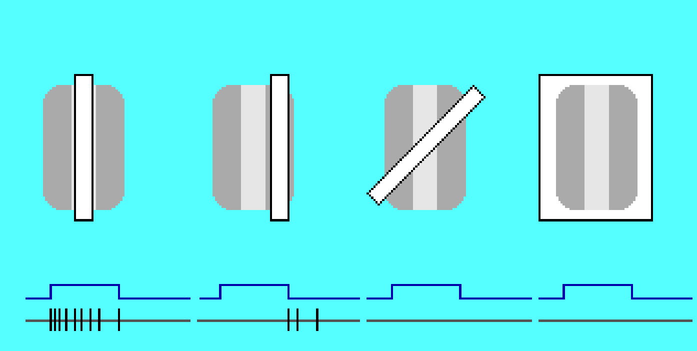


Figure 3.2 Simple cells are selective for orientation, size and position and respond best to a bar or an edge of correct orientation and size in correct position on retina(Hubel 1995).

Complex cells tend to have larger RFs than simple cells, but not much larger. A typical complex RF in the fovea of the macaque monkey would be about one-half degree by one-half degree.

Complex cells do not have clearly defined excitatory and inhibitory regions and thus respond to both light increment and decrement throughout their RFs. Complex cells are the most numerous in V1 (constituting roughly three-quarters of the population). Like simple cells, they respond only to appropriately oriented stimuli, but unlike simple cells, they are not sensitive to the position of the stimulus, as long as it falls somewhere inside the RF (see Fig. 3.3). By contrast, to evoke sustained responses from a simple cell, a stationary line must be critically oriented and critically positioned; a moving line evokes only a brief response at the moment it crosses a boundary from an inhibitory to an excitatory region or during the brief time it covers the excitatory region. Complex cells that do react to stationary slits, bars, or edges fire regardless of where the line is the RF, as long as the orientation is appropriate. Many complex cells are also direction-selective, in the sense that they respond only when the stimulus moves in one direction or when it moves in the opposite direction.

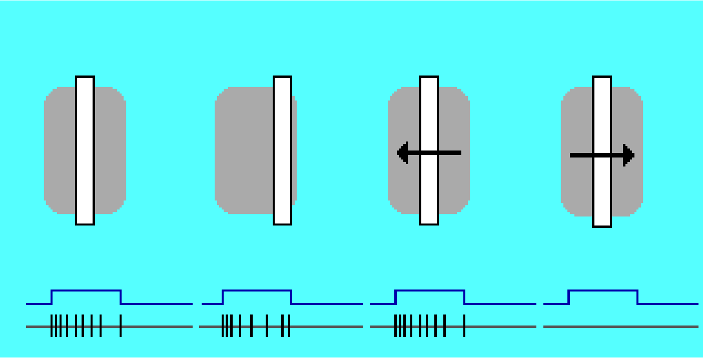


Figure 3.3 Complex cells respond best to rapidly flickering or moving stimuli, they have a preferred orientation and size as simple cells, but are not as specific for stimulus position (Hubel 1995).

The plausible and simplest scheme for a complex cell is that the complex cell receives input from many simple cells, all of whose fields have the same orientation but are spread out in overlapping fashion over the entire field of the complex cell, s shown in Fig. 3.4. If the connections from simple to complex cells are excitatory, then wherever a line falls in the field, some simple cells are activated; the complex cell will therefore be activated. If, on the other hand, a stimulus fills the entire RF, none of the simple cells will be activated, and the complex cell won't be activated.

The burst of impulses from a complex cell to the appearance of a stationary line that is not moving is generally brief even if the light is kept on; we say that the response adapts. When we move the line through the complex cell's RF, the sustained response may be the result of overcoming the adaptation by bringing in new simple cells one after the next.

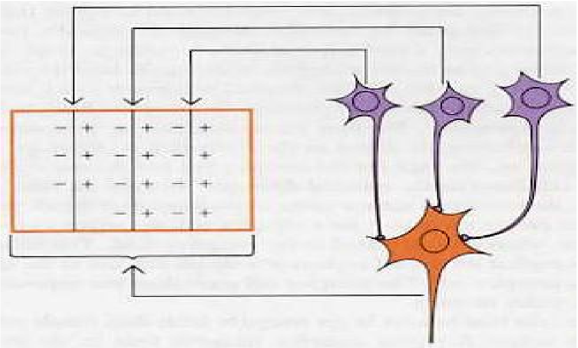


Figure 3.4 A large number of simple cells (only three are shown here) make excitatory synapses with a single complex cell. Each simple cell responds optimally to a vertically oriented edge with light to the right, and the RFs are scattered in overlapping fashion throughout the rectangle. An edge appearing anywhere within the rectangle evokes a response from a few simple cells, and this in turn evokes a response in the complex cell. Because there is adaptation at the synapses, only a moving stimulus will sustain a steady bombardment of the complex cell (Hubel 1995).

Simple and complex cells can work together to decompose the outlines of a visual image into short segments, the basis of simple and complex object recognition.

3.1.2 Redundancy Reduction

Horace Barlow has argued that redundancy provides knowledge. Redundancy in sensory input contains structural information about the environment. What is important for the perceptual system to detect is "suspicious coincidences," new statistical regularities in the sensory input that differ from the environment to which it has been adapted. Bars and edges, for example, are locations in the visual input at which there is phase alignment across multiple spatial scales, and therefore constitute a "suspicious coincidence" in Barlow's formulation (Barlow 1994). Learning mechanisms that encode the redundancy that is expected in the input and that remove it from the output, enable the system to more reliably detect these new regularities. Incoming sensory stimuli are automatically compared against the null hypothesis of statistical independence, and suspicious coincidences signaling a new causal factor can be more reliably detected. Learning such a transformation is equivalent to modeling the prior knowledge of the statistical dependencies in the input. Independent codes are advantageous for encoding complex objects that are characterized by higher-order combinations of features, because the prior probability of any particular higher-order combination is low.

Redundancy reduction has been discussed in relation to the visual system at several levels. A first-order redundancy is mean luminance. Adaptation mechanisms take advantage of this nonrandom feature by using it as an expected value, and expressing values relative to it. Contrast gain control, which has been observed in V1 (Heeger 1992), takes account of changes in the variance of the input signals. Principal component analysis (PCA) is a way of encoding second order dependencies in the input by rotating the axes to correspond to directions of maximum covariance. PCA provides a dimensionality-reduced code that separates the correlations in the input. Atick and Redlich (Atick and Redlich 1992) have argued for such compact, decorrelated representations as a general coding strategy for the visual system. Second-order statistics capture the amplitude spectrum of images but not the phase (Bell and Sejnowski 1997). The Fourier transform of the autocorrelation function of a signal is equal to its power spectrum (the square of the amplitude spectrum). The remaining information that is not captured by the autocorrelation function, the high order statistics, corresponds to the phase spectrum. Spatial phase contains the structural information in images that drives human recognition much more strongly than the amplitude spectrum (Oppenheim and Lim 1981; Piotrowski and Campbell 1982).

Independent component analysis (ICA) (Lee 1998) is a generalization of PCA that separates the higher-order dependencies in the input, in addition to the second-order dependencies. The difference between ICA and PCA is illustrated as follows. Consider a set of data points derived from two underlying distributions as shown in Fig. 3.5. PCA encodes second order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. PCA models the data as a multivariate Gaussian and would place an orthogonal set of axes such that the projections of the two distributions would be completely overlapping. ICA does not constrain the axes to be orthogonal, and attempts to place them in the directions of statistical dependencies in the data. Each weight vector in ICA attempts to encode a portion of the dependencies in the input, so that the dependencies between the elements of the output are removed. The projection of the two distributions onto the ICA axes would have less overlap, and the output distributions of the two weight vectors would be kurtotic (Bartlett and Sejnowski 1997).

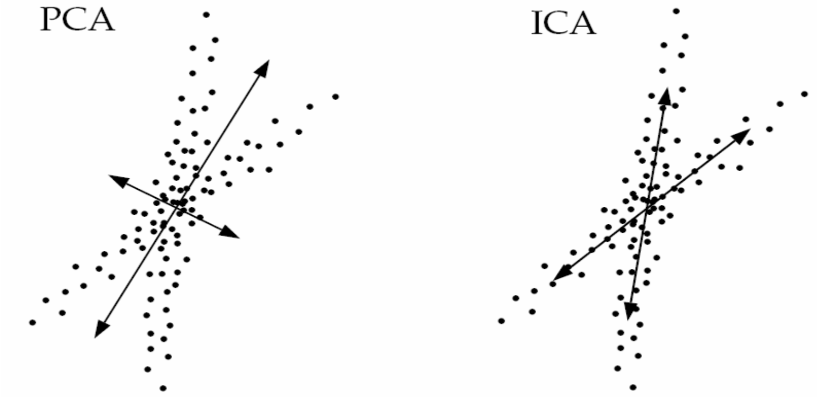


Figure 3.5 Example 2-D data distribution and corresponding principal component and independent component axes (Bartlett, Lades et al. 1998).

Bell and Sejnowski (Bell and Sejnowski 1995) recently developed an algorithm for separating the statistically independent components of a dataset through unsupervised learning. For linear ICA, one considers an ensemble of signals, each produced by an unknown linear superposition of unknown independent (elementary) signals. By presenting the ICA algorithm with a large number of examples of such signals, it is able to reconstruct the elementary signals; at least if the elementary signals have non-gaussian probability densities (i.e. the distribution of the strengths with which each elementary signal is present in a set of images is not a Gaussian). ICA on natural images produces RFs like those of simple cells (Bell and Sejnowski 1997; Hateren and Schaaf 1998).

3.2 Learning the Receptive Fields of V1 Simple Model Cells

3.2.1 Image Model

Image representations are often based on discrete linear transformations of the observed data. Consider a black-and-white image whose gray-scale value at the pixel indexed by x and y is denoted by $I(x, y)$. Many basic models in image processing express the image $I(x, y)$ as a linear superposition of some features or basis functions $a_i(x, y)$:

$$I(x,y)=\sum_{i=1}^{n}a_i(x,y)s_i \tag{3.1}$$

where the s_i are stochastic coefficients, different for each image $I(x,y)$. The image code is determined by the choice of basis functions a_i. The goal of efficient coding is to find a set of a_i that forms a complete code (i.e., spans the image space) and results in the coefficient values being as statistically independent as possible over an ensemble of natural images. Alternatively, we can just collect all the pixel values in a single vector $X=(x_1,x_2,......x_m)^T$, in which case we can express the representation as

$$\mathbf{X}=\mathrm{A}\mathbf{s} \tag{3.2}$$

The goal of a perceptual system is to linearly transform the images, X, with a matrix of filters, W, so that the resulting vector:

$$\mathbf{u}=\mathbf{W}\mathrm{x} \tag{3.3}$$

and recover the underlying causes, **s**, possibly in a different order and rescaled, as illustrated in Fig. 3.6. Representing by **P,** an arbitrary permutation matrix (all zero except for a single "one" in each row and each column), and by **S**, an arbitrary scaling matrix (non-zero entries only on the diagonal), such a system has converged when:

$$\mathbf{u}=\mathbf{W}\mathbf{A}\mathrm{s}=\mathbf{P}\mathbf{S}\mathrm{s} \tag{3.4}$$

The scaling and permuting of the causes are arbitrary, unknowable factors, so we will consider the causes to be defined such that **PS**=**I** (the identity matrix). Then the basis functions (columns of **A**) and the filters which recover the causes (rows of **W**) have the simple relation: **W=A**$^{-1}$. All that remains in defining an algorithm to learn **W** (and thus also **A**) is to decide what constitutes a "cause".

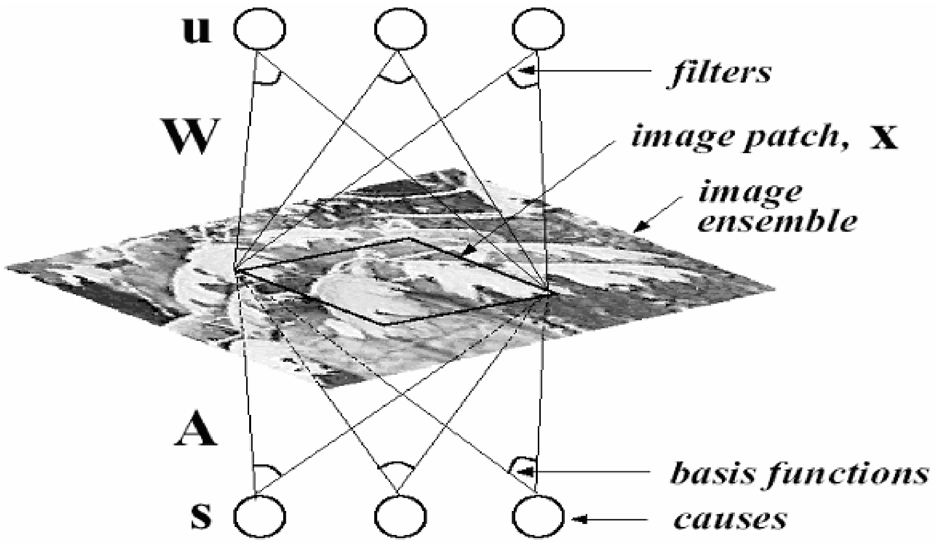


Figure 3.6 The Blind Linear Image Synthesis model. Each patch, X, of an image is viewed as a linear combination of several underlying basis functions, given by the matrix **A**, each associated with an element of an underlying vector of "causes", **s**, which are viewed as statistically independent "image sources". The causes are recovered (in a vector **u**) by a matrix of filters, W, more loosely "RFs" (Bell and Sejnowski 1997).

3.2.2 Independent Component Analysis (ICA)

Barlow proposed an organizational principle for unsupervised learning based on information theory and argued for minimum entropy coding as a general representational strategy. Minimum entropy, highly kurtotic codes, have low mutual information between the elements. This is because the joint entropy of a multidimensional code is defined as the sum of the individual entropies minus the mutual information between the elements. Since the joint entropy of the code stays constant, by minimizing the sum of the individual entropies, the mutual information term is also minimized. Another way to think of this is moving the redundancy from *between* the elements to redundancy *within* the distributions of the individual elements. The distributions of individual elements with minimum entropy are redundant in the sense that they almost always take on the same value (Barlow 1994).

In natural images, oriented lines and edges, especially curved, fractal-like edges, give rise to statistical dependencies that are of higher-order than linear pairwise correlations, and so it is important to consider these forms of structure as well when developing an efficient code. The goal here will be to find a linear coding strategy that is capable of reducing these higher-order forms of redundancy.

ICA is a generalization of PCA which separates the higher-order moments of the input, in addition to the second-order moments. It is an unsupervised learning rule that was derived from the principle of optimal information transfer through sigmoidal cells. Consider the case of a single input, x, and output, y, passed through a nonlinear squashing function, g, as illustrated in Fig. 3.7.

$$\begin{aligned} u &= wx + w_0 \\ y &= g(u) = \frac{1}{1+e^{-u}} \end{aligned} \tag{3.5}$$

The optimal weight w on x for maximizing information transfer is the one that best matches the probability density of x to the slope of the nonlinearity. The optimal w produces the flattest possible output density, which maximized the entropy of the output.

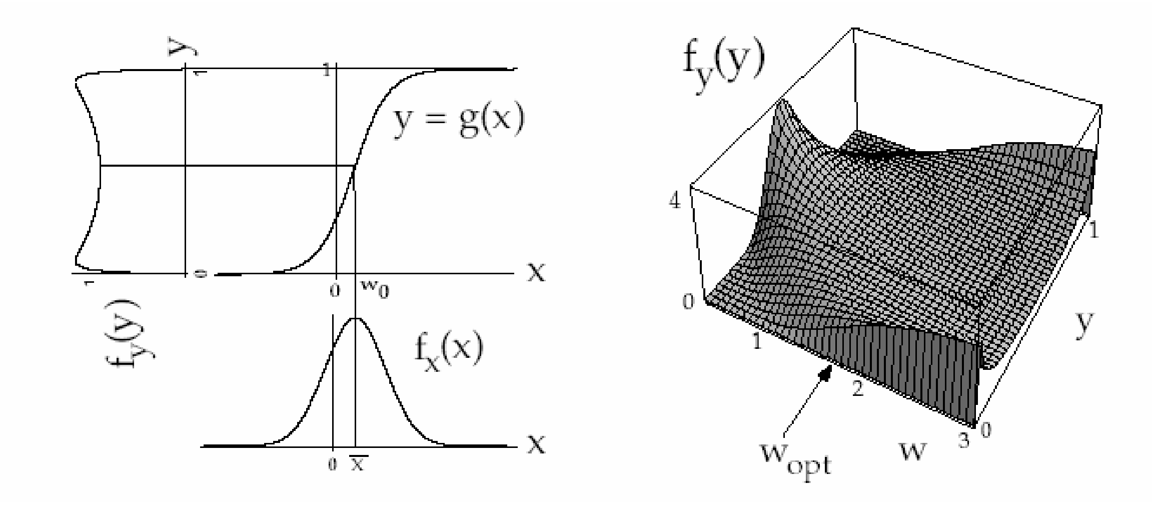


Figure 3.7 Optimal information flow in sigmoidal cells. The input x is passed through a nonlinear function, g(x). The information in the output density $f_y(y)$ depends on matching the mean and variance of $f_x(x)$ to the slope and threshold of g(x). Right: $f_y(y)$ is plotted for different values of the weights, w. The optimal weight, w_{opt} transmits the most information (Bell and Sejnowski 1995).

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

Considering the joint entropy at the outputs of a neural network is

$$H(y_1,...,y_N) = H(y_1) + \cdots + H(y_N) - I(y_1,...,y_N) \tag{3.6}$$

$H(y_i)$ is the marginal entropies of the outputs and $I(y_1,.....y_N)$ is their mutual information. Each marginal entropy can be written as

$$H(y_i) = -E[\log p(y_i)] \tag{3.7}$$

where $E\{\bullet\}$ denotes expected value. Let $y_i = g_i(u_i)$ with $g_i(u_i)$ is an invertible monotonic nonlinearity, the probability density of the output y_i can be described using the probability density of the estimated independent sources, u_i.

$$p(y_i) = \frac{p(u_i)}{\left|\frac{\partial y_i}{\partial u_i}\right|} \tag{3.8}$$

Substituting Eq. (3.8) into Eq. (3.6), gives

$$H(Y) = -E[\log \frac{p(u_1)}{\left|\frac{\partial y_1}{\partial u_1}\right|}] + \cdots + E[\log \frac{P(u_N)}{\left|\frac{\partial y_N}{\partial u_N}\right|}] - I(Y) \tag{3.9}$$

We can see that maximizing $H(y_1,...y_N)$ consists of maximizing the marginal entropies and minimizing the mutual information. When the form of the nonlinear transfer function g is the same as the cumulative density functions of the underlying independent components (up to a scaling and translation), i.e. eliminating all the marginal entropy terms to zero by setting $\frac{\partial y_i}{\partial u_i} = p(u_i)$, it can be shown that maximizing the mutual information between the input X and output Y also minimizes the mutual information between outputs

$$H(Y) = -I(Y) \tag{3.10}$$

The mutual information is always positive and is zero if and only if the components are independent. Now the direct maximization of the joint entropy between output components implies the minimization of the mutual information which makes the output component independents.

Because $H(Y) = -E\{\log P(Y)\}$ and $P(Y) = \frac{P(X)}{|J(X)|}$ with $|J(X)|$ is the absolute value of the Jacobian of the transformation from X to$Y$, we get

$$H(Y) = E\{\log|J(X)|\} - E\{\log P(X)\} \tag{3.11}$$

and

$$J(X) = \det\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_N}{\partial x_1} & \cdots & \frac{\partial y_N}{\partial x_N} \end{bmatrix} \tag{3.12}$$

with $\frac{\partial y_i}{\partial x_j} = w_{ij}\frac{\partial y_i}{\partial u_j}$ when u=Wx. The maximum of the joint entropy$H(Y)$ can be found by deriving $H(Y)$ with respect to$W$, i.e. computing the gradient of $H(Y)$. The term $-E\{\log P(X)\}$ in Eq. (3.11) does not depend on the parameter$W$, therefore, it is equal to zero. Now considering a training set of the data X so that the stochastic learning rule can be approximated without the expectation term

$$\frac{\partial H(Y)}{\partial W} = \frac{\partial(E\{\log|J(X)|\}}{\partial W} \approx \frac{\partial \log|J(X)|}{\partial W} \tag{3.13}$$

Since there are no connections between the outputs of the cell, the partial derivative $\frac{\partial y_i}{\partial u_j}$ is non zero for $i = j$ only, the Jacobian can be rewritten as

$$J(X) = \det(W)\prod_{i=1}^{N}\left|\frac{\partial y_i}{\partial u_i}\right| \tag{3.14}$$

From Eq. (3.14) in Eq. (3.13) it follows

$$\begin{aligned} \frac{\partial H(Y)}{\partial W} &= \frac{\partial}{\partial W}\log\left(|\det(W)|\prod_{i=1}^{N}\left|\frac{\partial y_i}{\partial u_i}\right|\right) \\ &= \frac{\partial}{\partial W}\log|\det(W)| + \frac{\partial}{\partial W}\sum_{i=1}^{N}\log\left|\tfrac{\partial y_i}{\partial u_i}\right| \\ &= \frac{\partial}{\partial W}\log|\det(W)| + \sum_{i=1}^{N}\frac{\partial}{\partial W}\log\left|\tfrac{\partial y_i}{\partial u_i}\right| \end{aligned} \tag{3.15}$$

The first term in Eq. (3.15) is

$$\frac{\partial}{\partial W}\log\left|\det(W)\right| = \frac{(adjW)^T}{\det W} = (W^T)^{-1} \tag{3.16}$$

because the determinant of W can be expressed as

$$\det(W) = \sum_{j=1}^{N} w_{ij}cof(w_{ij}) \tag{3.17}$$

The second term in Eq. (3.15) can be further computed as

$$\frac{\partial}{\partial w_{ij}}\sum_{i=1}^{N}\log\left|\tfrac{\partial y_i}{\partial u_i}\right| = \frac{1}{\frac{\partial y_i}{\partial u_i}}\frac{\partial^2 y_i}{\partial u_i^2}x_j \tag{3.18}$$

let
$$\hat{y}_i = \frac{1}{\frac{\partial y_i}{\partial u_i}}\frac{\partial^2 y_i}{\partial u_i^2} = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i} \tag{3.19}$$

Now the first and second term in Eq. (3.15) are computed and the leaning rule is

$$\frac{\partial H(Y)}{\partial W} = (W^T)^{-1} + \hat{y}X^T \tag{3.20}$$

where $\hat{y} = \left[\hat{y}_1, \cdots, \hat{y}_N\right]^T$.

A much more efficient way to maximize the joint entropy is to follow the "natural" gradient (Amari, Cichocki et al. 1996). The natural gradient rescales the entropy gradient by post-multiplying the entropy gradient by W^TW giving

$$\Delta W \alpha \frac{\partial H(Y)}{\partial W}W^TW = (I + \hat{y}u^T)W \tag{3.21}$$

This rule has two advantages over Eq. (3.20) of avoiding the matrix inverse and of converging several orders of magnitude more quickly, for data X ,that is not pre-whitened.

3.2.3 Experiments and Results

Nine natural images involving trees, leafs and so on are taken for this experiment (see Fig.3.8). 50,000 patches of size 12×12 pixels from those images were randomly selected. Care must then be taken to avoid border effects. Each sample is a column in the training set $X(144 \times 5000)$. Followed Bell and Sejnowski's work, the training set was "sphered" by subtracting the mean and multiplying by two times the local symmetrical (zero-phase)

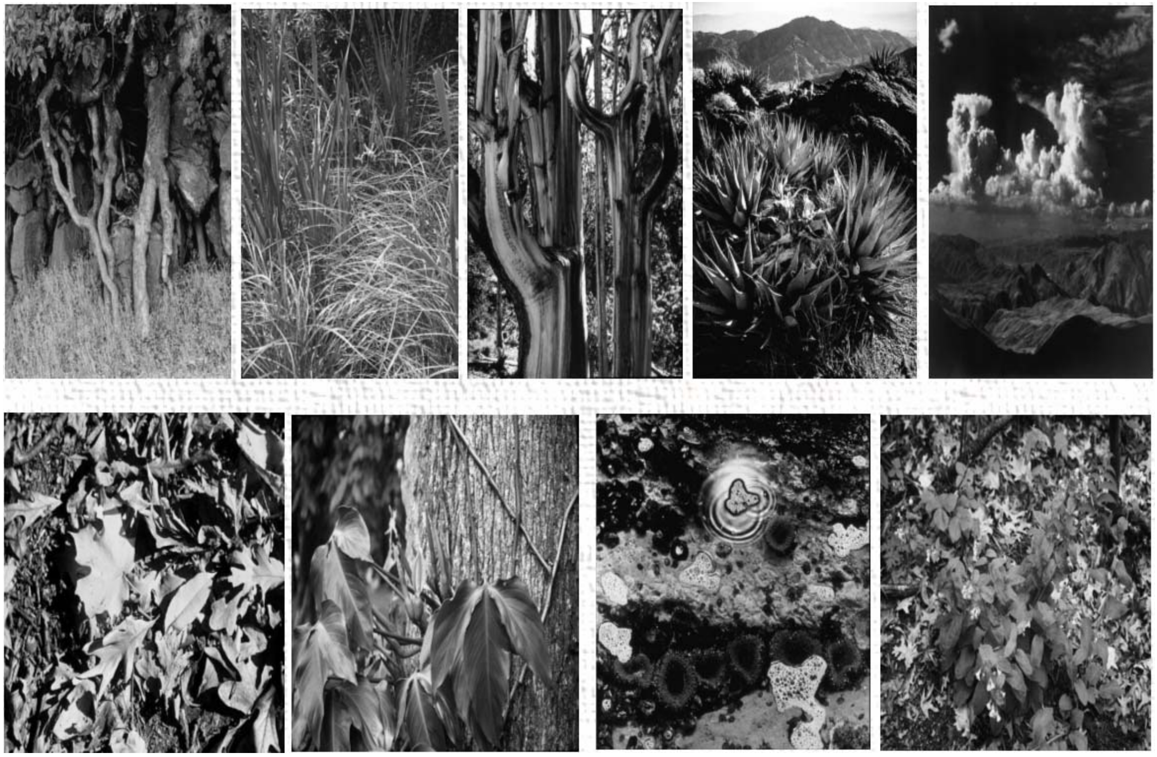


Figure 3.8 Nine natural images used for training the RF of V1 model cells.

whitening filter of $W_Z = \left\langle XX^T \right\rangle^{-\frac{1}{2}}$:

$$\{X\} \leftarrow 2Wz(\{X\} - \langle X \rangle) \qquad (3.22)$$

This removes both first and second order statistics from the data and makes the covariance matrix of X equal to 4I. By pre-whitening X in this way, we can ensure that subsequent transformation, u=Wx, to be learned should approximate an orthonormal matrix (rotation without scaling), roughly satisfying the relation $W^T W = I$, where I is identity matrix. The RFs (12×12) of simple model cells (M-cells), i.e. the rows of W, were learned using the Eq. (3.21) with the following parameters: block size of 50, the learning rate is changeable and 60 passes through the data (i.e. epoch size is 60). A momentum (=0.1) was used as well. The learned W had a dimension of 144×144. Figure 3.9 shows the complete set of learned filters. This process took 2.5 hours running MATLAB on a Gateway PC with 256MB RAM. The full ICA transforms from the raw image was calculated as the product of the sphering matrix and the learnt matrix:

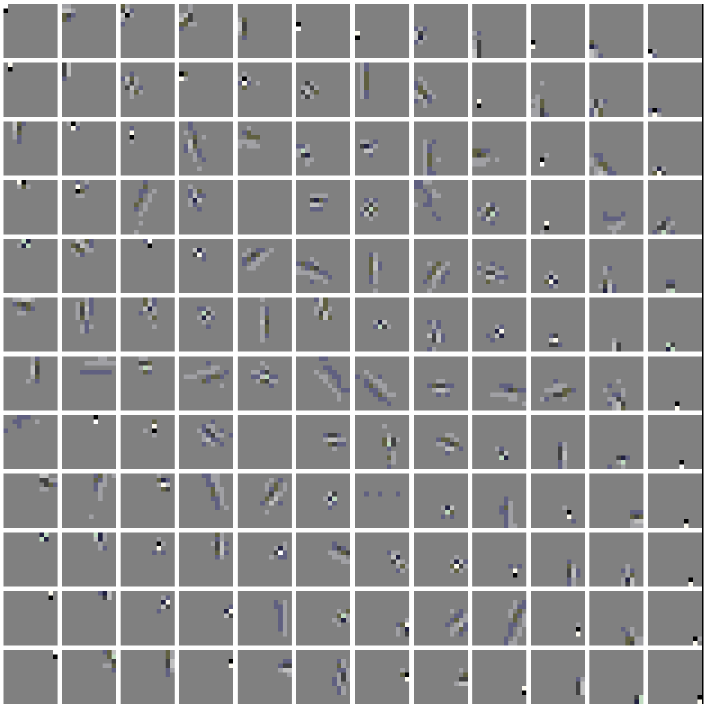


Figure 3.9 ICA on natural images produces 144 filters like the RFs (12×12) of V1 simple M-cells. Each filter is a row of the matrix W. The diagonal filters are longer than the vertical and horizontal due to the bias induced by having square, rather than circular, RFs.

$W_I = W * W_Z$. Also, the kurtosis of the 144 filters, which measures the 4th moment relative to the variance squared, was calculated by

$$K = \frac{\sum_i (u_i - \overline{u})^4}{(\sum_i (u_i - \overline{u})^2)^2} - 3 \tag{3.23}$$

The kurtotic ICA distribution has a peak and a longer tail, indicating that it is sparser than those of other methods, such as PCA. The basis function functions **A**, was calculated as W_I^{-1}. A PCA matrix (see Fig. 3.10), W_P, was calculated as

$$W_P = D^{-\frac{1}{2}} E^T \tag{3.24}$$

where D is the diagonal matrix of eigenvalues, and the columns of E is the eigenvectors of the covariance matrix.

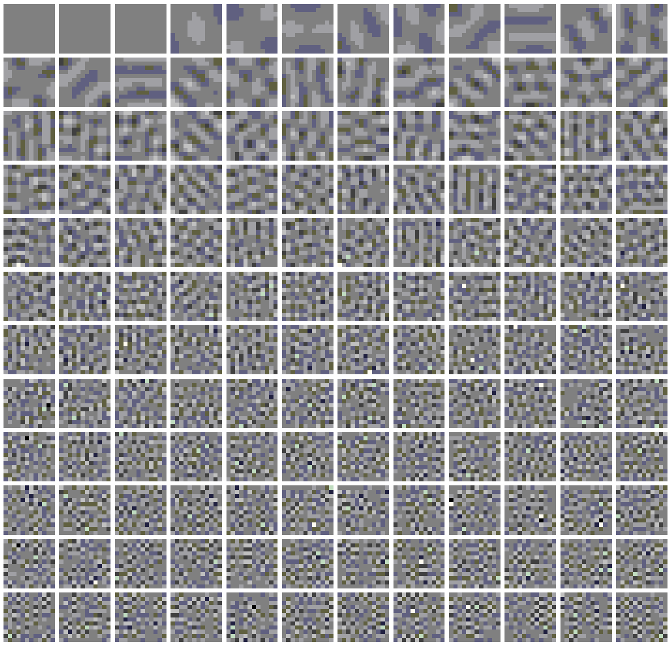


Figure 3.10 Principal components calculated on 12×12 image patches extracted from nine natural images. The full set of 144 components is shown, ordered by their variance. The oriented structure of the first few principal components does not arise as a result of the oriented structures in natural images, but rather because these functions are composed of a small number of low frequency components (the lowest spatial frequencies account for the greatest part of the variance in natural scenes).

The filters resulting from training on natural scenes are displayed in Fig. 3.9 and Fig. 3.10. The PCA filters are spatially global and ordered in frequency. The ICA filters are localized and mostly oriented. The ICA filters consist of one DC filter, 3 filters close to the DC filters and 72 oriented filters. 46 of them were diagonal, 12 are vertical and 14 horizontal. There are also 48 localized checkerboard patterns.

It is known that, the RFs of V1 simple biological cells are localized in space and time, have band-pass characteristics in spatial and temporal frequency domains, are oriented, and often sensitive to the direction of motion of a stimulus. This sort of properties encourages the notion that the purpose of the cells in V1 is to construct economical representation of the images. Similar to RFs of simple cells, ICA on natural images produces the localized edge detectors, which are semi-local and sensitive to both phase (locality) and frequency information, just as in transforms involving oriented Gabor functions (Daugman 1985) or wavelets.

Edges are the first level of invariance in images, detected by liner filters alone. Further levels of invariance (shifting, rotating, scaling, and lighting) clearly exist with natural objects in natural settings. These further levels may be extractable by multi-layer networks with non-linear mechanisms.

3.3 Learning the Receptive Fields of V1 Complex Model Cells

An essential behavior of animals is the visual recognition of objects that are important for their survival. Human activity, for instance, relies heavily on the classification or identification of a large variety of visual objects (Logothetis and Sheinberg 1996). One of the major problems which must be solved by a visual system for object recognition is the building of a representation of visual information which allows recognition to occur relatively independent of size, contrast, spatial frequency, position on the retina, and angle of view, etc (Ullman, Vidal-Naquet et al. 2002). This requires that features extracted by the visual pathway create a rather complete representation of the current sensory scene using the principle of sparse coding, which means that at any one time only a small selection of all the units is active, yet this small number firing in combination suffices to represent the scene effectively.

Hubel and Wiesel proposed a model in which V1 simple cells with neighboring RFs feed into the complex cell with the same receptive-field orientation and roughly the same positions, thereby endowing that complex cell with phase and shift invariant features. Consistent with this approach, we introduce a complex-cell layer to pool the energies (squared outputs) of groups of simple cells that are near-by on the topographic grid (see Fig. 3.11). By extending the basic ICA to maximize the sparseness of local activations, the two-layer network can model complex cell properties.

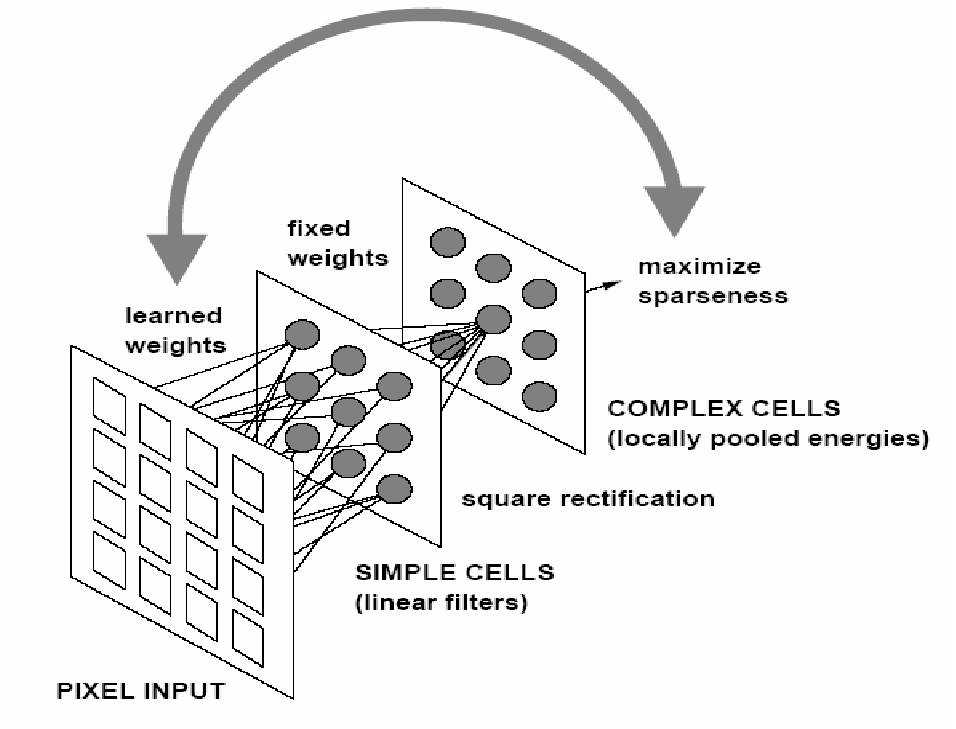


Figure 3.11 Illustration of the two-layer network. The first layer consists of simple cells or garbor filters. The second layer is a complex cell layer where the energies of simple cell outputs are pooled locally (Inspired by (Hubel 1995) .

3.3.1 Independent Feature Subspace

In classic ICA, the independent components have no particular order, or other relationships. In practical applications of ICA, however, one can observe clear violations of the independence assumption. It is possible to find, for example, couples of estimated independent components such that they are clearly dependent on each other. This remaining dependence is often informative and can be further analyzed.

Kohonen developed the principle of invariant-feature subspaces as an abstract approach to representing features with some invariances. This principle states that one may consider an invariant feature as a linear subspace in a feature space. The value of the invariant, higher-order feature is given by (the square of) the norm of the projection of the given data point on that subspace, which is typically spanned by lower-order features (Kohonen 1996). A feature subspace, as any linear subspace, can always be represented by a set of orthogonal basis vectors, say $w_i(x,y), i=1,...,M$, where M is the dimension of the subspace. Then the value $F(I)$ of the feature F with input vector $I(x,y)$ is given by $F(I)=\sum_{i=1}^{m}<w_i,I>^2$, and $s_i=<w_i,I>$. In fact, this is equivalent to computing the distance between the input vector $I(x,y)$ and a general linear combination of the weights (filters) $w_i(x,y)$ of the feature subspace. A graphical depiction of feature subspace is given in Fig. 3.12.

In contrast to ordinary ICA, the components s_i in multidimensional independent component analysis (Cardoso 1998) are not assumed to be all mutually independent. Instead, it is assumed that the s_i can be divided into couples, triplets or in general m-tuples, such that s_i inside a given m-tuple may be dependent on each other, but dependencies between different m-tuples are not allowed. When embedding invariant-feature subspaces in multidimensional ICA analysis, the logarithm of the likelihood L of the K observed image patches $I_k(x,y), k=1,...K$ is

$$\begin{aligned} L &= \log l(I_k(x,y),k=1,...,K;w_i(x,y),i=1,...,M) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\log p(\sum_{i\in S_j}<w_i,I_k>^2)+K\log\left|\det W\right| \end{aligned} \tag{3.25}$$

where $S_j, j=1,...,J$ denote j-th subspace; $p(\sum_{i\in S_j}s_i^2)=p_j(s_i,i\in S_j)$ gives the probability density (pdf) inside the j-th subspace of s_i, and W is a matrix containing the filters $w_i(x,y)$ as its columns. Prewhitening the image patches $I_k(x,y)$ allows us to consider the $w_i(x,y)$ to be orthonormal, which implies that $\log\left|\det W\right|$ is zero, then

$$\Delta W \propto \frac{\partial L}{\partial W}=I(x,y)<w_i,I>g(\sum_{r\in S_{j(i)}}<w_r,I>^2) \tag{3.26}$$

where j(i) is the index of the subspace to which w_i belongs, and $g={p'}\!/\!{p}$ is a nonlinear function.

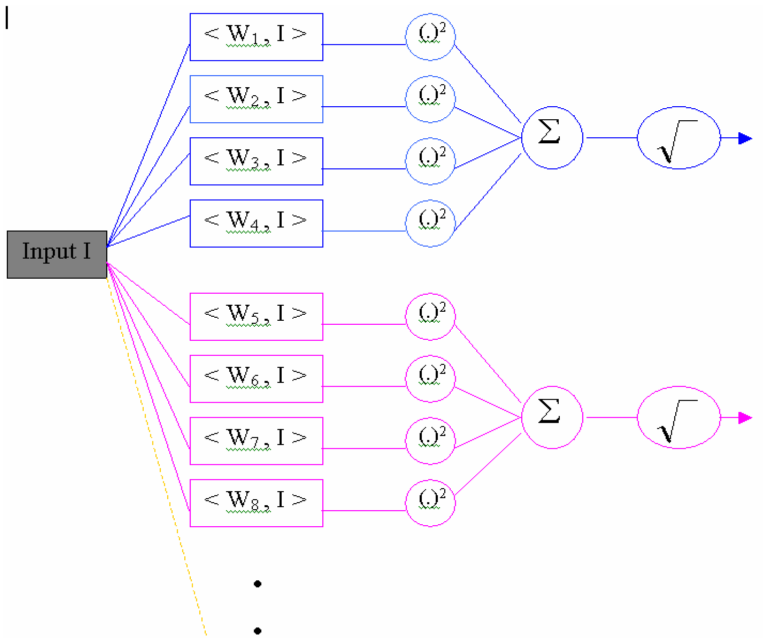


Figure 3.12 Graphical depiction of the feature subspaces. First, dot products of the input data with a set of basis vectors are taken. Here, we have two subspaces with four basis vectors in each. The dot products are then squared, and the sums are taken inside the feature subspaces. Thus we obtain the (squares of the) norms of the projections on the feature subspaces, which are considered as the responses of the subspaces. Square roots may be taken for normalization (inspired by (Hyvarinen and Hoyer 2000)).

Since the norm of the projection of visual data on practically any subspace has a supergaussian distribution, we need to choose the probability density p in the model to be sparse, so we could use the following pdf

$$\log p(\sum_{i\in S_j} s_i^{\,2}) = -\alpha\left[\sum_{i\in S_j} s_i^{\,2}\right]^{1/2} + \beta \qquad (3.27)$$

and then

$$g(u) = p'/p = -\frac{1}{2}\alpha u^{-1/2} \qquad (3.28)$$

After training the weights in the V1 layer by the above algorithm, we compute the response (CC_j) of the j-th complex cell as

$$CC_j = \sqrt{\sum_{i\in S_j}(COV(I(x,y),rf_i))^2} \qquad (3.29)$$

where $COV(.)$ denotes the convolution of input image $I(x,y)$ with rf_i, where rf_i is the RF corresponding to weights w_i.

3.3.2 Experiments and Results

V1 complex cells share the properties of simple cells but have two distinguishing properties of phase invariance and (limited) shift invariance. Extending ICA by combining the principles of invariant-feature subspace and multi-dimensional ICA, the features similar to those found in complex cells emerged from maximizing the independence/sparseness between the different feature subspaces (Hyvarinen and Hoyer 2000; Hyvarinen and Hoyer 2000).

Followed the work by Hyvärinen and Hoyer, the image patches (16×16 pixels) for training the RFs of complex M-cells were randomly sampled from twelve monochrome natural images involving trees, leaves, animals, and so on. The training set $X=\{I_k, k=1,...,K\}$was pre-whitening by: (a) subtracting the mean gray-scale value from the data, this removes the first order statistics; (b) the dimension of the data was then reduced by the PCA with the largest variances; the PCA filter is $W_P = D^{-1/2}E^T$, where we have $EDE^{-1} = \left\langle XX^T \right\rangle$, and D is the diagonal matrix of eigenvalues, and the columns of$E$is the eigenvectors of the covariance matrix. Using the random initial value for W, the likelihood of 50,000 such observations was maximized under the constraint of orthonormality of the filters in the whitened space. By the learning rule in Eq. (3.26), we learned 40 complex M-cells (subspaces) with the subspace dimension 4 (see Fig. 3.13). Next, we computed the responses of complex M-cells by Eq. (3.29). This process took 3 hours running MATLAB on a Dell Precision Station (530 MT). Figure 3.14 shows the response of one of these complex M-cells by testing with a grayscale dog image. It can be seen that the basis vectors associated with a single complex M-cell all have approximately the same orientation and frequency. Their locations are not identical, but close to each other. The phases differ considerably. Intuitively, the responses of V1 M-cells are very sparse.

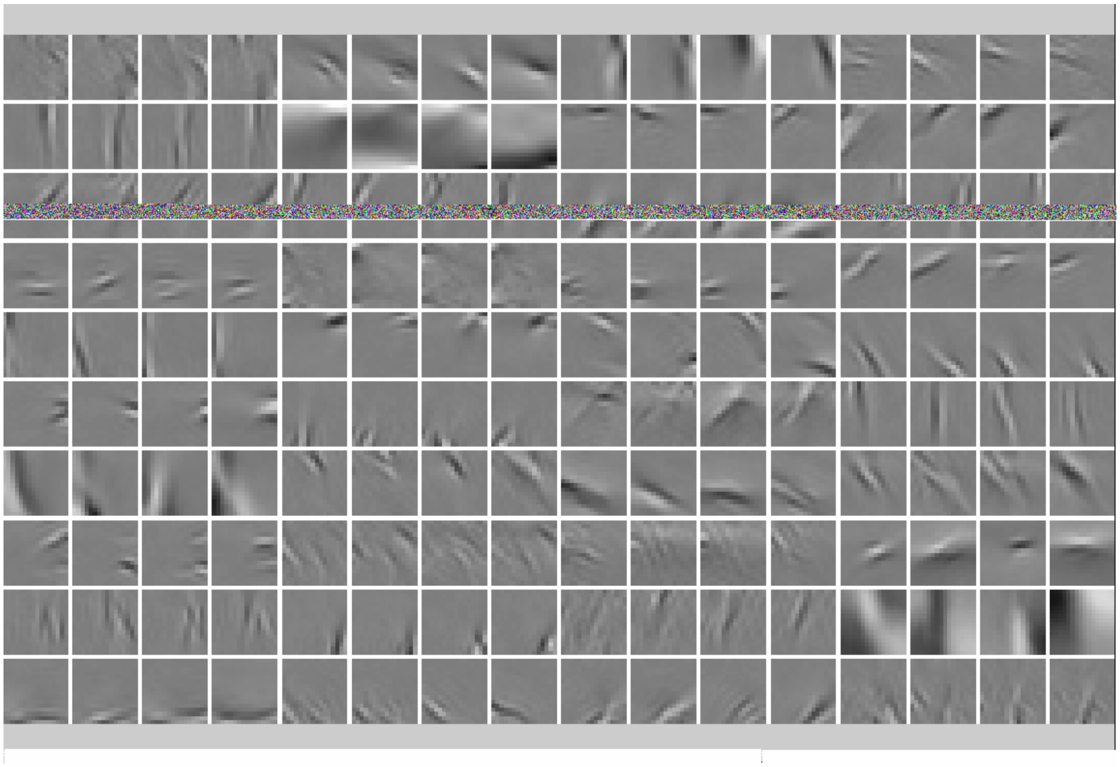


Figure 3.13 The learned RFs of 40 model complex cells. The basis vectors associated with a single complex M-cell all have approximately the same orientation and frequency. Their locations are not identical, but close to each other. The phases differ considerably. Intuitively, the responses of V1 M-cells are very sparse.

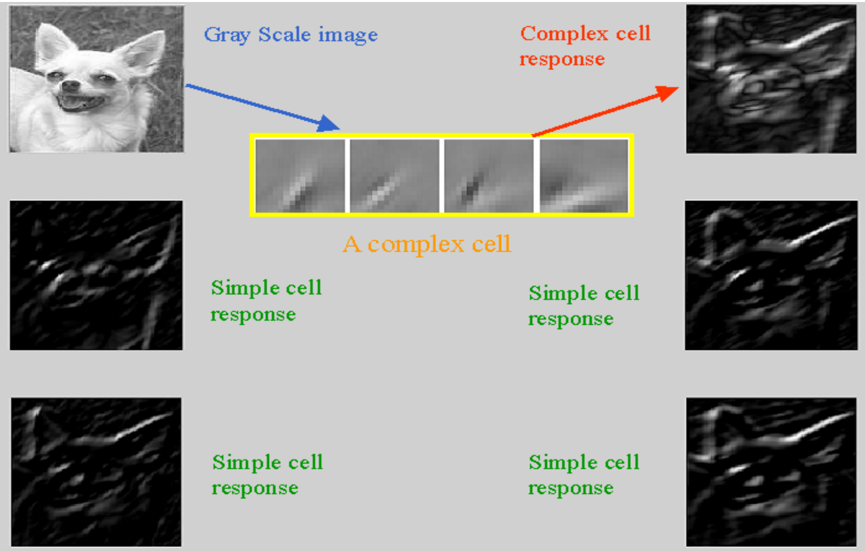


Figure 3.14 Give an example for the learned complex M-cell and its response. For comparison, we also give the responses of four basis vectors that constitute the RFs of simple M-cells.

3.4 Leaning the Receptive Fields of V2 Model Cells

V2 is at least as large as V1 and shows alternating thick, thin and pale stripes. V2 is strongly reciprocally connected with V1. Cells in V2 and V1 are distinguished by the sizes of their RFs. Most cells in V2 are orientation-selective, though perhaps slightly less sharply so than cells in V1; their preferred spatial frequencies are a little lower. In most respects cells in V1 and V2 are not remarkably different, having broadly the same mix of directional selectivity and similar distributions of chromatic preferences. However, end stopping seems to be more prevalent in V2, particularly in the pale stripes.

End-stopped cells are responsive to lines of specific length. They have very large RFs that may combine complex cell signals. An ordinary simple or complex cell usually shows length summation: the longer the stimulus line, the better the response, until the line is as long as the RF; making the line still longer has no effect. For an end-stopped cell, lengthening the line improves the response up to some limit, but exceeding that limit in one or both directions results in a weaker response, as illustrated in Fig. 3.15. We can view end-stopped cells (act as angle detectors) as sensitive to corners, to curvature, or to sudden breaks in lines.

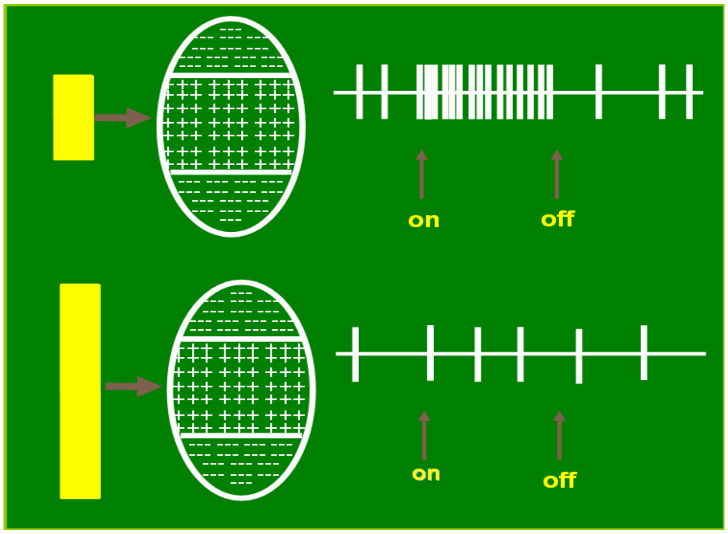


Figure 3.15 For an end-stopped cell, lengthening the line improves the response up to some limit, but exceeding that limit in one or both directions results in a weaker response.

3.4.1 Non-negative Sparse Coding

We showed how the classical RFs of simple cells in V1 can be understood in the framework of sparse coding. In a two-layer network, we further showed how complex cell properties emerged by maximizing the independence, or equivalently, the sparseness of the norms of the projections to feature subspaces (the energies of groups of simple M-cells). An important question then is how to extend these models to account for response properties of cells higher in the processing hierarchy. Perhaps the most straightforward approach is to add a layer on the top of the complex cell model as illustrated in Fig. 3.16. This would amount to assuming a model where the activities of complex M-cells are not independent, but rather can be described as a linear combination of some higher-order independent components by

$$X = \sum_{i=1}^{n} a_i s_i + n \tag{3.30}$$

where n is Gaussian noise. Each complex M-cell activation pattern gives on data vector X , with each x_j representing the firing rate of one M-cell. Each s_i represents the response of one higher-order M-cells, whose "RFs" is closely related to the corresponding a_i .

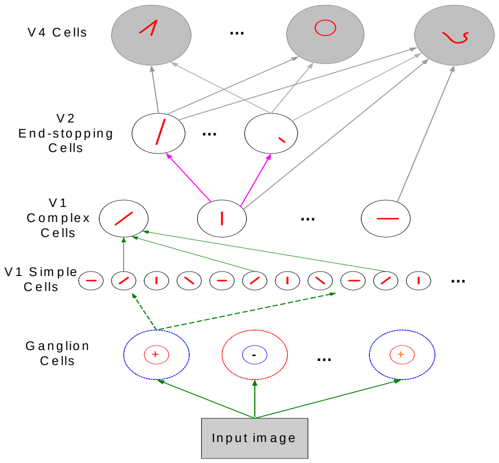


Figure 3.16 A hierarchical network architecture (up to but not including V4) inspired by the visual ventral pathway. Complex cell (model cell) responses are calculated in a feedforward manner and these responses are subsequently analyzed by a higher-order sparse coding layer (i.e. V2 layer) in the network.

Here we assume s_i is non-negative and sparse, following the response properties of complex cells. The goal is to find basis patterns a_i such that typical input patterns X can be described accurately using only a few significantly active V2 M-cells.

In the language of probability theory, we wish to match as closely as possible the distribution of observed patterns arising from the imaging model in Eq. (3.30), $P(X/A)$, to the actual distribution of patterns observed in nature, $P^*(X)$. To assess how well this match is, we take the Kullback-Leibler (KL) divergence between the two distributions

$$KL = \int P^*(X)\log\frac{P^*(X)}{P(X/A)}dX \tag{3.31}$$

because $P^*(X)$ is fixed, minimizing KL amounts to maximizing $<\log P(X/A)>$. Since

$$<\log P(X/A)> = \int P^*(X)\log P(X/A)dX \tag{3.32}$$

the goal of learning will be to find a set of basis A^* that maximize the average log-likelihood of the observed patterns under a sparse, statistically independent prior, such that $A^* = \arg\max_A <\log P(X/A)>$. We can express the objective in an energy function framework by defining

$$E(X,S/A) = -\log P(X/S,A)P(S) \tag{3.33}$$

Then, we have
$$E(X,A/S) = \sum_n \left[\left\| X^{(n)} - AS^{(n)} \right\|^2 + \lambda \sum_i s_i^{(n)} \right] \tag{3.34}$$

where $S^{(n)}$ is the vector containing the latent variables $s_i^{(n)}$ corresponding to the n-th observed vector $X^{(n)}$, and the constant λ defines the tradeoff between representation error and sparseness. The objective (E) was minimized by standard gradient descent with respect to $S^{(n)}$ in the short timescale and with respect to A under a longer timescale (Olshausen and Field 1997). In Eq. (3.34), the first term computes the reconstruction error; the second term incurs a penalty on the coefficient activities, which encourage sparse representation.

Learned by the above algorithm, the weights (RFs of M-cells) in V2 layer are selective for contour length in addition to being tuned to position and orientation, and exhibit end-stopping properties. It has been proposed that contour feature extraction is the ultimate purpose of end-stopping.

3.4.2 Experiments and Results

We first used the method in section 3.3.2 to compute 200 complex M-cell responses (CC) from five natural images, where $CC = \{CC_{i\,j}, i = 1,...,5; j = 1,...,40\}$ in the V1 layer. Then, the training set $X = \{x_n, n = 1,..., N\}$ for the V2 layer was obtained by randomly extracting 24×24 patches (recall that the RF of a V2 cell is typically 2~3 times larger than that of a V1 cell) from 50 complex M-cell responses among CCs. Using 20,000 such patches, we trained the weights in the V2 layer using the methods described above. Combining the sparse coding and non-negative constraint, after 40 iterations, the learned 288 weights/RFs of V2 M-cells are shown in Fig. 3.17. This process took 6 hours by running MATLAB on the same computer mentioned above. Visually, the basis patterns are in different positions, different orientations, and different lengths. Moreover, for characterizing the learned V2 M-cell RFs (24×24), we approximated them in the parameter space as done in (Hoyer and Hyvarinen 2002). The main results are shown in Fig. 3.18, which shows a richer tuning of orientation and length than what has been reported before. This kind of length tuning, or the property of the end-stopped cell, is very interesting for visual features representation. As illustrated in Fig. 3.19, the necessity for different length basis patterns comes from the fact that long basis patterns simply cannot code short (or curved) contours and short basis patterns are inefficient at representing long, straight contours. Figure 3.20 shows the activity patterns of V2 cells or the source vectors s_i in the image model, Eq. (3.30). It demonstrates that V2 cells produce a sparse-distributed representation for any given input, because the number of V2 cells responding to each particular instance of the input is minimized, while the dimensionality of the representation is maintained.

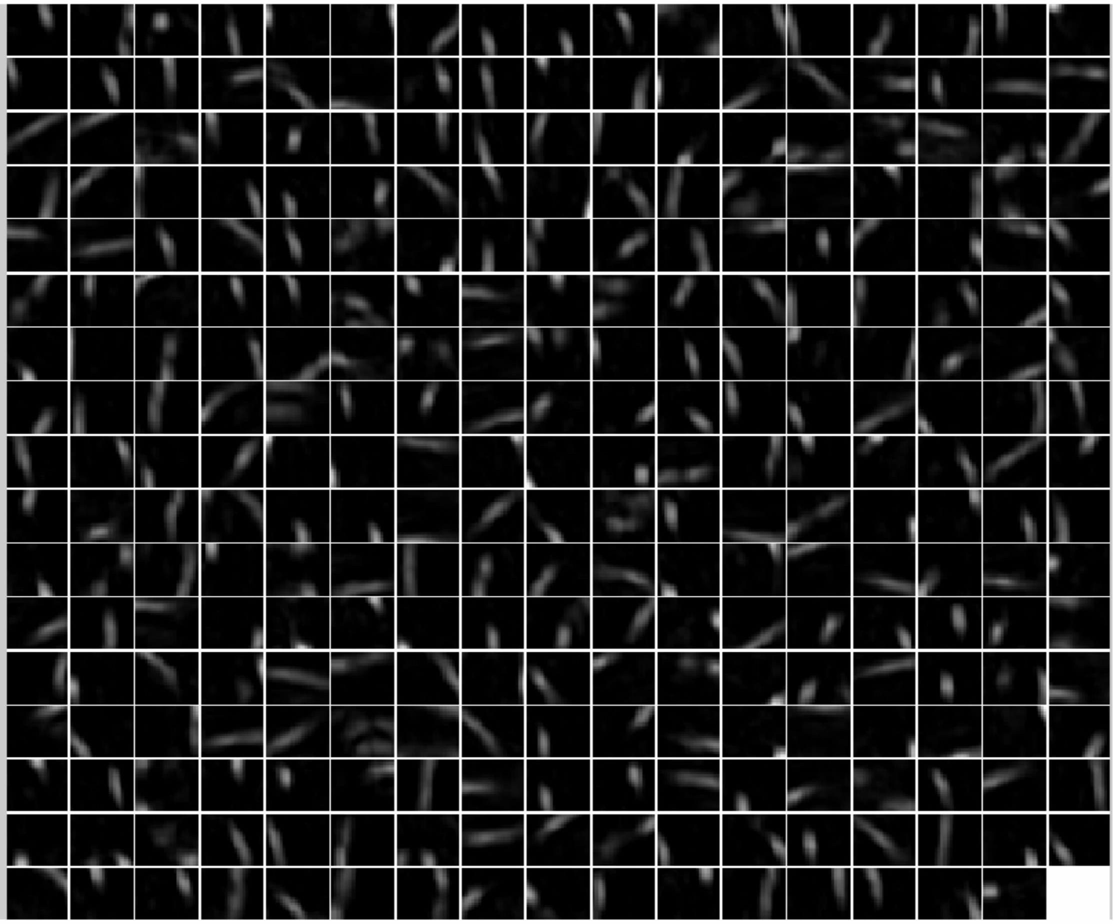


Figure 3.17 The 288 RFs (24×24) learned in the V2 layer. They are in different position, different orientations and different lengths.

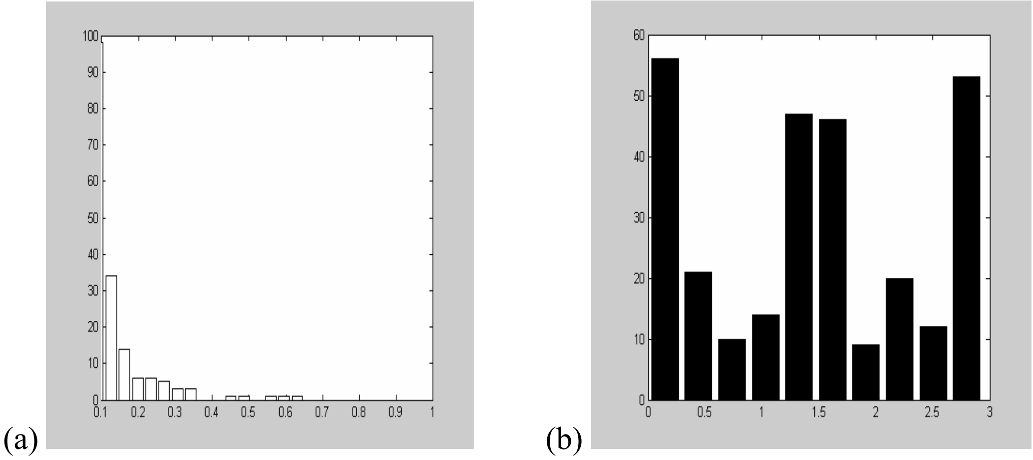


Figure 3.18 (a) Distribution of the RFs lengths in V2 layer, which are normalized by the width of the sampling window; (b) Distribution of the RF orientations (from 0° to 180°: label 0~3 in the horizontal axis) in V2 layer.

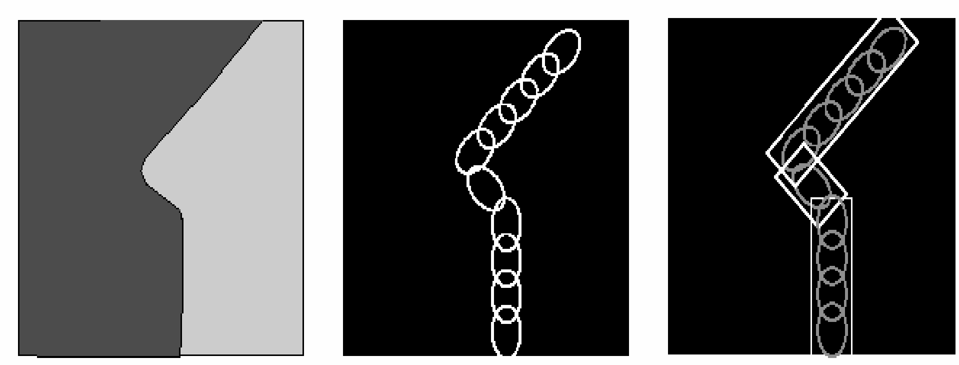


Figure 3.19 The necessity for different length basis patterns come from the fact that long basis patterns simply can not code short (or curved) contours, and short basis patterns are inefficient at representing long, straight contours

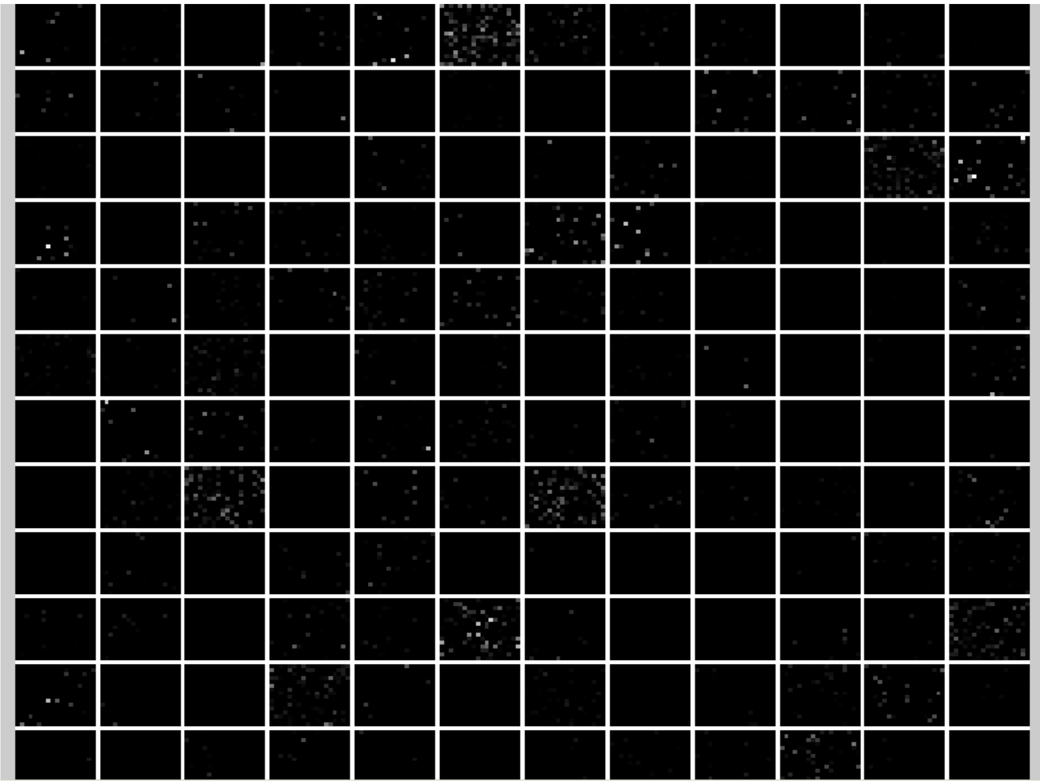


Figure 3.20 The activity patterns of V2 end-stopped M-cells. It demonstrates that V2 M-cells produce the sparse-distributed representation of any given input, because the number of V2 M-cells responding to the particular instance of the input is minimized, while the dimensionality of the representation is maintained.

3.4.3 Discussion

In this chapter, we developed a two-layer neural network with sparse coding constraint, which extracted low level features (such as edges, orientations, spatial frequencies, and contours) for further processing in the ventral pathway, such as part-based shape representation in cortex V4. This sparse representation provides an intrinsic low power and fault-tolerant computing substrate to sensorimotor control systems. By unsupervised learning algorithms, the learned visual models made the sensorimotor control systems to automatically adapt to uncertain and novel environment. We also showed that in such a model, V2 M-cell RFs develop end-stopping properties. According to Hubel and Wiesel, the optimal stimulus for an end-stopped cell is a line that only extends for a certain distance. For a cell that responds to edges and is end-stopped at one end only, a corner is ideal; for a cell that responds to slits or black bars and is stopped at both ends, the optimum stimulus is a short white or black line or a line that curves so that it is appropriate in the activating region and inappropriate (different by 20 to 30 degrees or more) in flanking regions. We can thus view end-stopped cells as sensitive to corners, to curvature, or to sudden breaks in a line. These contours are crucial for shape representation in cortex V4 (Gallant, Braun et al. 1993; Wilkinson, James et al. 2000; Pasupathy and Connor 2001), and subsequently they are very important for object representation and recognition in IT.

Our approach is related to Hoyer's contour coding network (Hoyer and Hyvarinen 2002). However, Hoyer computed the complex M-cell responses using a simple energy model; therefore the RFs in his V1 layer are fixed, or pre-calculated. In contrast, our approach uses the end-to-end learned RFs (recall: the RFs of our V1 layer were learned by extended ICA), and thus represents the natural image sparsely and sufficiently (see Fig. 3.21). Also, the property of the RFs and their sizes in our architecture are richer and more in-line with the diversity that is common in biology. Repeating Hoyer's experiments using 100,000 image patches and 100 iterations took 2 days on the same computer mentioned above, the selective resulting basis patterns are shown in Fig. 7. Practically, using the responses of V2 M-cells in our architecture, we have trained the V4 layer and obtained some interesting results, such as object parts. However, the responses of V2 cells produced by Hoyer's model are too weak to be further used in higher layers.

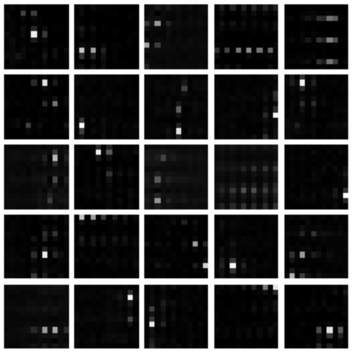


Figure 3.21 A selective set of basis function learned by Hoyer's network.

Our study is also related to the predictive coding model of (Rao and Ballard 1999), in which, the feedback connections from a higher- to a lower- order visual area carry predictions of lower-level neural activities. The feedforward connections carry the residual errors between the predictions and the actual lower-level activities. They proposed that end-stopping cell stopped responding when the stimulus length was increased because then it could be predicted and there were no residual errors. Contrastingly, training monkeys over weeks and months usually increases the incidence of neurons with highly specific RFs (Kobatake, Wang et al. 1998), rather than decreasing their number as would be expected if cortex were implementing a predictive coding strategy.

Chapter 4

Parts-based Shape Representation

4.1 Theories of Shape Representation

Shape representation is a major challenge in machine perception, and is the basis for vision (Biederman 1987; Pentland 1987). It is also important in related fields, such as robotics and the understanding of the human vision system. Many researchers (Binford 1971; Marr and Nishihara 1978; Marr 1982; Nevatia 1982; Pentland 1987; Rao 1988; Palmer 1999) have discussed the requirements for shape representation. A good shape representation should be rich (i.e. data preserving), to allow discrimination between similar shapes. It should be stable so that small changes in the shape result in small changes in the representation. It also should be invariant to changes in the viewing conditions (e.g. lighting, rotation, size and position etc.). Moreover, shape is by far the most important and complex of all visually perceivable properties(Palmer 1999). Its significance derives from the fact that it is the most informative visible property, because shape allows a perceiver to predict more facts about an object than from any other property. It is complex because it is a combination of many different attributes.

How people perceive shapes is certainly among the most difficult problems in visual perception. To date, the major theories that have been proposed as solutions are templates, Fourier analysis, feature lists and structural descriptions (i.e. parts-based shape representation).

Templates are always the starting point for discussions of shape representation, for the simple reason that they are the most obvious mechanism for performing this conversion. We know with a fair degree of certainty that the visual system makes use of templates to represent very simple “shapes”, such as line and edge detector in V1 cortex. However, as a general theory of shape representation, template theory has a number of major weakness. Many visual factors have little, if any, impact on perceived shape yet strongly influence the correlation between a template and an input image. For example, a green square on a yellow ground is seen as having the same shape as a blue square on a red ground or any other combination of colors. However, a template representing a green square on a yellow background will not match an otherwise identical template for a blue square on a red ground because the underlying color elements are different. Comparing the template representations of two objects with the same shape does not generally result in a good fit if they differ only with respect to one of the similarity transformations: translation, rotation, dilation, reflection, and their various combinations. In other words, template representations are not invariant with respect to these visual factors and transformations, whereas shape perception is.

A second representational scheme that has been proposed for shape representation is based on the spatial frequency content of images, or their Fourier spectrum (Ginsburg 1971; Ginsburg 1986). The Fourier approach has several noteworthy factors in its favor. First, it not only is consistent with, but was actually derived from, a prominent theory of low-level vision. Another attractive feature of the Fourier approach to shape is its formal mathematical status. Fourier analysis is a well-known and fully developed technique that can be applied to image processing with all the mathematical knowledge that has accumulated since Fourier proved his important theorem in 1822. This mathematical rigor stands in contrast to virtually all other theories of shape representation, whose formal foundations might charitably be called “uncertain”. A more tangible advantage of frequency based methods is that relative power spectra appear to solve the problem of shape equivalence over the similarity transformations. Despite these significant strengths, the Fourier analysis has a number of drawbacks that seriously undermine their plausibility as a viable basis for shape representation. One of the most important is that a global Fourier analysis represents an entire, interpreted image rather than individual objects. This means that the representation of the shape of a figure will be inextricably intertwined with that of the background on which it appears. This is a fatal flaw, because the same object on different backgrounds can have dramatically different spectra.

Wavelets and Gabor approaches are strongly associated with Fourier analysis and partially overcome its shortcomings (Unser 1994; Wickerhauser 1994). Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. The wavelet transform is often compared with the Fourier transform, where signals are represented as s sum of sinuosids. The main difference is that wavelets are localized in scale and in time where the standard Fourier transform is only localized in frequency. A Gabor mask is a sinusoidal waveform which is spatially localised by modulation with a Gaussian envelope. Mathematically, the family of Gabor convolutions is a spatially localised modification of a Fourier analysis.

For several decades, the most popular class of shape representation was feature lists, i.e. a symbolic description consisting of a simple set of attributes. According to this view, an object's perceived shape is defined by the set of its spatial features, and the degree of similarity between two shapes can be measured by the degree of correspondence between the two feature sets. The features that have typically been proposed for representing shape are of two types: global properties, such as symmetry, closeness, and connectedness, and local properties, such as containing a straight line, a curved line, or an acute angle. The reason for the popularity of feature representation is that they overcome some of the simple and basic objections that limit template and Fourier approaches. However, feature theories do have several important weaknesses. A difficult problem with feature list theories is that it is often unclear how to determine computationally whether a given object has the features that are proposed to make up its shape representation. Assuming that the computational difficulties of extracting features can be overcome, there is still the very difficult problem of specifying what the proper features for a shape representation are. In general, an underlying problem of feature list representation is that describing the shape of an object adequately, often requires specifying not only its component parts and global properties, but also the spatial relations that hold among its parts. These part relational features do not fit with the conception of feature lists as a simple set of independently specifiable attributes, because they inherently depend on the features that represent the parts. This suggests a more complex form of representation that directly encodes relational information among explicitly represented parts.

Structural descriptions are representations that contain explicit information about parts and about relations between parts. They are usually depicted as networks in which nodes represent the whole object and its various parts, while labeled links between nodes represent the relations between parts. Such a representation of relational information is far more economical than corresponding representations in feature theories because one does not need a different feature for all possible combinations of relations among parts. Rather, the network structure allows for the relations to be separate entities that connect the nodes to each other. Structural descriptions overcome many of the difficulties with templates, Fourier spectra, and feature lists. They are invariant over different visual factors, such as luminance, texture and motion. They can account for the effects of spatial transformation on shape perception by absorbing them within object-centered reference frames. Compared with other shape processing theories, they are better adapted to the real-world difficulties of three-dimensional viewpoint transformation, partial occlusion and plastic deformation. Moreover, physiological results in higher level extrastriate cortex seem more compatible with structural description (Boussaoud, Desimone et al. 1991; Pasupathy and Connor 2001; Larsson, Amunts et al. 2002; Pasupathy and Connor 2002). At the most general level, there are two ways to go about dividing an object into parts, one relating to object boundaries and the other to solid volumes. The solid or volumetric primitive approach requires a set of simple, indivisible shapes that constitute the complete set of the most basic parts. More complex objects are then analyzed as configuration of these primitive parts. The boundary rules approach does not require any fixed set of primitive shapes but works directly on the whole object by dividing it into whatever parts result from the application of the boundary rules. Obviously, the key question of structural descriptions is how to extract a set of parts belonging to an object to achieve good and efficient shape description.

4.1.1 Object Parts

Indeed, shape depends critically on the part structure of objects and how its various parts are related to one another in terms of their relative positions, relative orientations, relative size and so forth. However, psychological findings also indicate important influences of global structure on local/part perception. A series of experiments by Palmer (Palmer 1980; Palmer and Bucher 1981; Palmer 1989) demonstrated that the perception of local spatial orientation is influenced by more global orientational structure. Other evidence (Reicher 1969; Weisstein and Harris 1974; Pomerantz, Sager et al. 1977) that global properties are important in human perception comes from experiments in which discrimination of parts is found to be superior when they are embedded within meaningful or well-structured wholes.

A "part" is a restricted portion of an object that has semi-autonomous, object-like status in visual perception. A human body, for example, is perceived as being composed of a head, a torso, two arms, and two legs; a standard wooden desk chair is perceived as consisting of a seat, a back, and four legs; a tree is perceived as consisting of a trunk, numerous limbs, and masses of individual leaves. In addition to these parts, object perceptions include the spatial relation among the parts. We also have the distinct impression that the parts we perceive are stable and non-arbitrary. We further believe that other people perceive them in pretty much the same way we do. Normal observers, for example, never see the lower half of a human head and the upper half of a human torso as a single natural body part, nor do they see the seat and one leg of a chair as a single natural chair part. Convincing evidence for the assertion that we perceive most objects as composed of parts comes from a number of sources. One of the most obvious kinds of evidence for the perceptual reality of objects parts comes from the language we have for talking about them. The perception of parts is so important and ubiquitous that we have separate words to refer to the salient parts of many familiar objects. Phenomenal experience (Palmer 1977; Hoffman and Richards 1984) also strongly suggests that perceiving complex objects gives rise to the spontaneous perception of parts. Further evidence of parts-based shape representation comes from recent neurophysiologic and neuropsychological studies (Versavel, Orban et al. 1990; Kobatake and Tanaka 1994; Hubel 1995; Gallant, Connor et al. 1996; Heeley and Buchanan-Smith 1996; Fahle 1997; Wilson, Wilkinson et al. 1997; Pasupathy and Connor 2001).

4.2 Parts-based Shape Representation in V4

Visual shape information is processed in the ventral cortical pathway (or so called "what" pathway), which hierarchically runs from V1 to V2, V4, and finally into various subregions of inferotemporal cortex (IT) (Ungerleider and Mishkin 1982; Felleman and Van Essen 1991). While response properties of the cells in V1 and V2 are comparatively well understood, the non-linear behaviors of cells in higher visual areas, such as V4 and IT, have made it difficult to determine the cortical computational mechanisms for increasing shape complexity. Several different theories have been proposed regarding response properties of cells in area V4. Although originally thought to be primarily involved in color vision (Zeki 1975; Zeki 1983), V4 is now recognized to be critical for shape representation. (Desimone, Schein et al. 1985; Desimone and Schein 1987) showed that many V4 cells exhibit length, width, orientation, direction of motion and spatial frequency selectivity, as in V1 and V2. (Kobatake and Tanaka 1994) found that some V4 cells respond better to complex shapes than to simple bar stimuli. (Gallant, Braun et al. 1993; Gallant, Connor et al. 1996) showed that the population of V4 cells is selective for polar, hyperbolic and Cartesian gratings. (Wilkinson, James et al. 2000) used functional magnetic resonance (fMRI) to demonstrate that the human V4 is more strongly activated by concentric and radial patterns than by conventional sinusoidal gratings. To further investigate the role of V4 in processing shape information, (Pasupathy and Connor 2001) studied 109 V4 cells that appeared to have complex shape response properties based on preliminary tests and suggested a parts-based representation of complex shapes in V4, where each part is a boundary pattern defined by curvature and position relative to the rest of the object.

4.3 Parts Decomposition

For decades, various theories and models have been proposed to explain how the shape of objects and their parts might be represented within the human visual system. However, current theories and models do not capture the astonishing power, versatility, and subtlety of human shape perception. How people perceive shape is certainly among the most difficult problems in visual perception, so difficult that no satisfactory solution has yet been proposed. Fortunately, significant progress has been made with new findings in visual physiology and psychology.

4.3.1 Model Structure

Inspired by the visual ventral pathway, we present here a shape representation model (see Fig. 4.1), which is different from previous work (Becker and Plumbley 1996; Burl, Weber et al. 1998; Nelson and Selinger 1998; Walker, Cootes et al. 1999; Fiser and Aslin 2001; Li, Hou et al. 2001; Edelman, Hiles et al. 2002; Edelman, Intrator et al. 2002) by integrating sparse representation and unsupervised learning properties into a hierarchical network. Shape representation is assumed to be a hierarchical process, with progressively more complex features at each stage. Local orientation of edges is considered to be the primary shape feature at early stages, based on the prevalence of orientation tuning in areas V1 and V2. The final representation involves a structural description based on parts and the relative positions of parts, similar with those reported by Pasupathy and Connor in V4 areas. Our model combines unsupervised learning methodology in the feedforward stream with sparse constraints to represent object shapes by parts.

In the hierarchical network, the receptive fields of V1 model cells (M-cells) were pre-trained by an independent feature subspace mechanism (Hyvarinen and Hoyer 2000). The output of the V1 layer is computed by convolving the receptive fields of V1 M-cells with an input image (Yang and Jabri 2003). Taking the output of the V1 layer as input, the synaptic weights onto V4 model cells (M-cells) are first trained by an unsupervised learning algorithm, modified non-negative matrix factorization (NMF), which is distinguished from other methods by its use of non-negativity constraints. These constraints lead to the sparseness of basis images (the receptive fields of V4 M-cells), because they allow only additive, not subtractive, combinations, therefore reducing the redundancy among the data (Field 1994; Foldiak and Young 1995; Lee and Seung 1999). Then a “winner-take–all” mechanism with lateral inhibition is used, by which, V4 M-cells compete with each other to win the selection according to the input. Finally, some of the synaptic weights of V4 M-cells are inhibitive. The synaptic weights remaining active represent the learnt receptive fields of V4 M-cells.

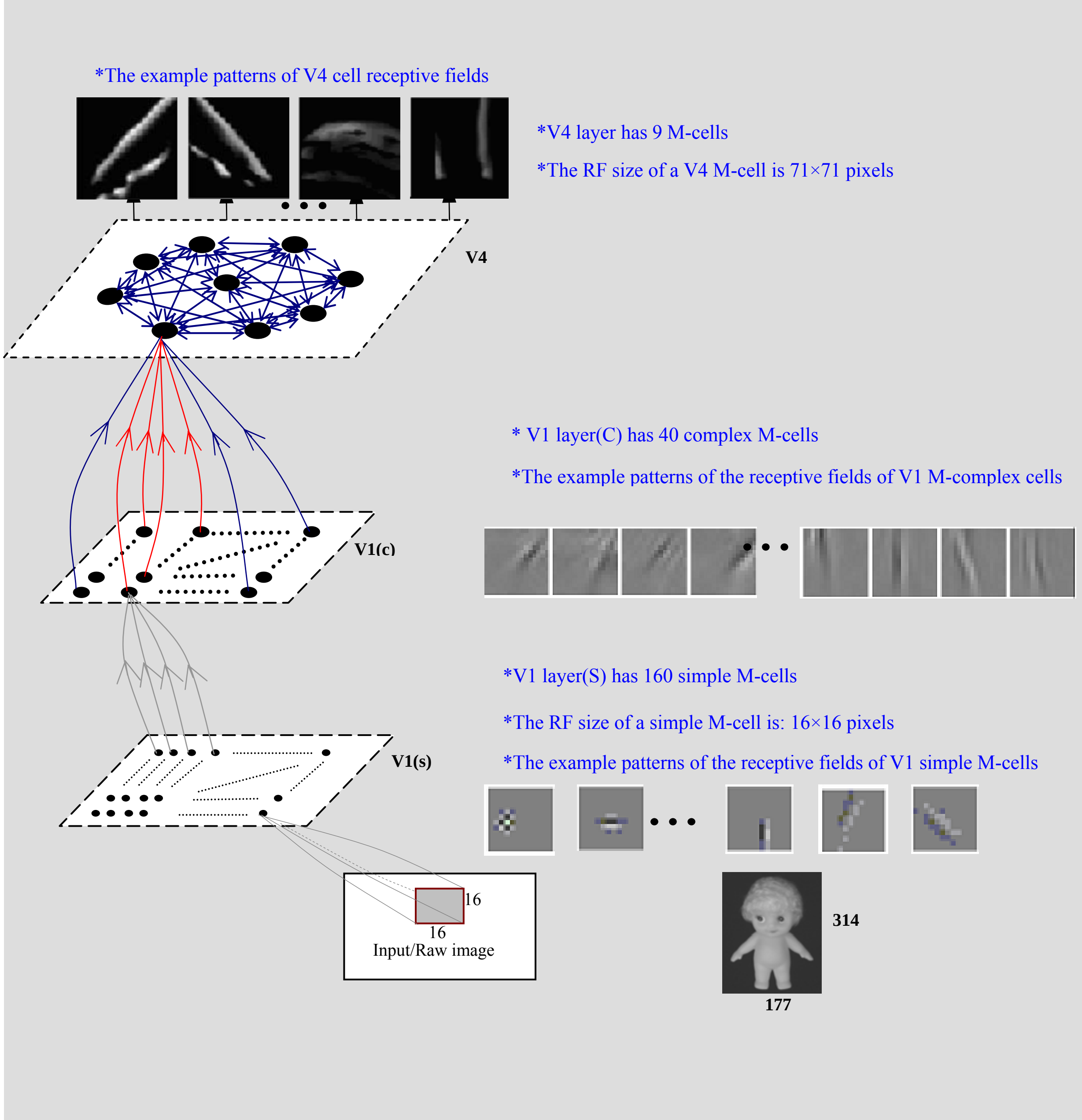

*The example patterns of V4 cell receptive fields
*V4 layer has 9 M-cells
*The RF size of a V4 M-cell is 71×71 pixels
V4
* V1 layer(C) has 40 complex M-cells
*The example patterns of the receptive fields of V1 M-complex cells
V1(c)
*V1 layer(S) has 160 simple M-cells
*The RF size of a simple M-cell is: 16×16 pixels
*The example patterns of the receptive fields of V1 simple M-cells
V1(s)
16
16
Input/Raw image
314
177

Figure 4.1 The visual processing pathway model, showing the structure of the V4 model. Here, the synaptic weights in the V1 layer were pre-learned (in gray).

4.3.2 Learning Algorithm for Part Decomposition

When non-negative constraints are used in the V4 layer, parts-based shape representations emerged because of two properties: the firing rates of cells are never negative and the synaptic strengths do not change sign. Given a non-negative matrix X , where X contains the response patterns C_j of V1 complex M-cells as its columns, we try to find non-negative matrix factors A and S, such that: $X \approx AS$. Each column of A is called a basis image. The columns of S correspond to coefficients by which the response pattern of a V1 complex M-cell is represented with a linear combination of basis images. To achieve this goal, we first define a cost function that quantifies the quality of the approximation. Such a cost function can be constructed using some measure of distance between two non-negative matrices U and V. One useful measure (Lee and Seung 2001) is

$$D(U \| V) = \sum_{ij} (U_{ij} \log \frac{U_{ij}}{V_{ij}} - U_{ij} + V_{ij}) \tag{4.1}$$

This is lower bounded by zero, and vanishes if and only if $U = V$. Because it is not symmetric in U and V, we will refer it as the "divergence" of U from V. It reduces to the Kullback-Leibler divergence, or relative entropy, when $\sum_{ij} U_{ij} = \sum_{ij} V_{ij} = 1$, so that U and V can be regarded as normalized probability distributions. Minimizing the divergence of X and AS as: $\min\limits_{A,S} D(X \| AS)$ with respect to A and S, and subject to non-negative constraints $A, S \geq 0$, we can find the local minimum of the objective function:

$$D(X \| AS) = \sum_{i,j} (x_{ij} \log \frac{x_{ij}}{\sum_{i,j} a_{ik} s_{kj}} - x_{ij} + \sum_{i,j} a_{ik} s_{kj}) \tag{4.2}$$

Further imposing the sparse constraints on the objective function, we modified the NMF as: (1) first minimize the number of basis images required to represent X by $\min \sum_i a_{ik}^2$, which implies that a basis image should not be further decomposed into more components; (2) then reduce the redundancy between different basis images by maximizing the orthogonality among basis images; and (3) by $\max \sum_k v_{kk}$ with $V = SS^T$, only basis images that provide the most important information are retained. Finally, the new learning rules are derived as:

$$S_{kj} \leftarrow \sqrt{S_{kj} \sum_i A_{ik} \frac{X_{ij}}{(AS)_{ij}}} \tag{4.3}$$

$$A_{ik} \leftarrow \frac{A_{ik} \sum_j \frac{X_{ij}}{(AS)_{ij}} S_{kj}}{\sum_j S_{kj}} \tag{4.4}$$

$$A_{ik} \leftarrow \frac{A_{ik}}{\sum_i A_{ik}} \tag{4.5}$$

The update rules preserve the non-negativity of A and S, and also constrain the columns of A to sum to unity. In the meanwhile, V4 M-cells are fully interconnected by lateral link weights following the inverted Mexican hat profile (Edelman, Intrator et al. 2002). With the objective function converging to a small value, V4 M-cells compete with each other by:

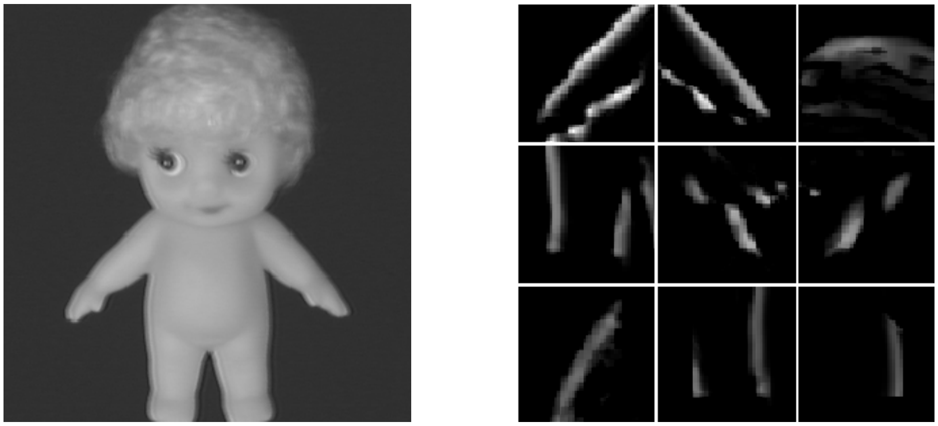
$$a_i = a_i - \beta a^* w_{mn} \tag{4.6}$$

and $$w_{mn} = G(d^*_{mn}, K\delta) - G(d^*{}_{mn}, \delta)$$

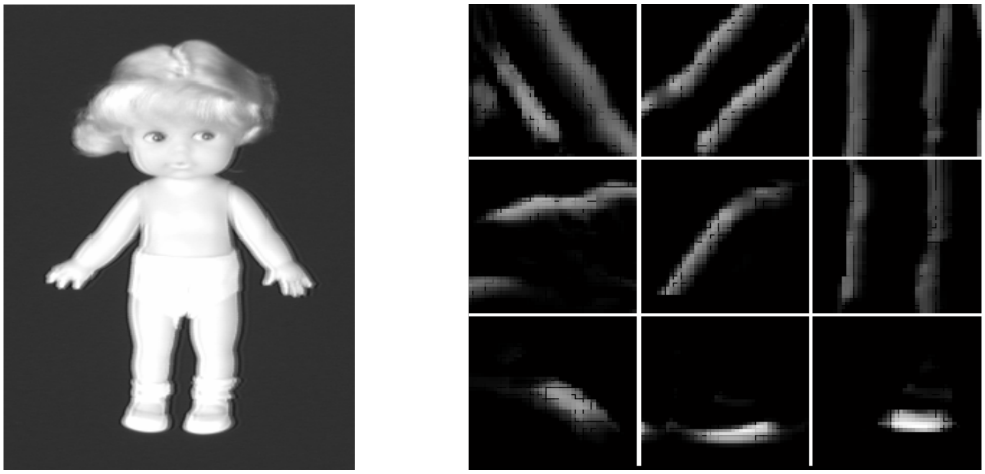
where $1 \le i \le N$ and N is the number of M-cells in V4 layer; a^* is the basis image (i.e. a column vector in A) corresponding to the winner V4 M-cell; w_{mn} is the strength of the lateral connection between M-cells m and n; $G(d,\delta)$ is the value of a Gaussian with δ width.

4.3.3 Experiments and Results

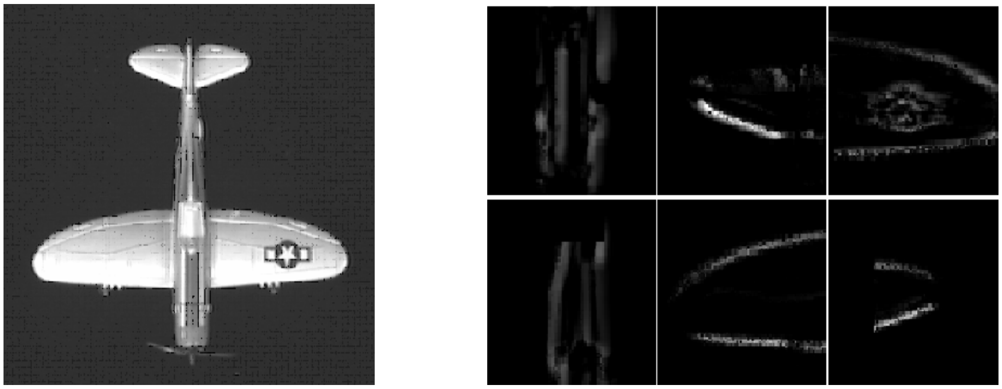
Gray images for part decomposition experiments were downloaded from ftp://ftp.cs.rochester.edu/pub/u/. The images contained objects of different sizes. The objects used here were chosen to be different in that they were easy for people to distinguish on the basis of shape. For an input image, the mean grayscale value was first subtracted so as to remove the first order statistics from the data, and then the grayscale intensities were normalized to the range $[0,1]$. In the V4 layer, the training data $X = \{CC_j, j = 1,...,40\}$ were obtained by using Eq. (3.29) to compute the response patterns of V1 complex cells, i.e. convolving the whitened image with pre-learned receptive fields (RFs) of V1 complex M-cells. Consistent with the visual system topography, the RF sizes of cells in the ventral pathway become larger by a factor of approximately 2.5 with succeeding stage and the receptive field sizes of cells in the IT (inferotemporal cortex) area can include the entire central visual field, we choose the receptive field size of V4 M-cells in our model as $RF_4\_SIZE = \sqrt{\frac{l_I^{\,2}}{2.5^2}}$ when l_I is the small dimension size of an input image (width×length). Starting with random initial conditions for A and S, the unsupervised learning algorithm usually converged after 45 iterations. This process took around 8 minutes running Matlab on a Dell Precision Station (530 MT, 2GHZ and 4GB). With different image sizes, the computation time varied by 1~2 minutes. Fig. 4.2 shows some examples of part decomposition by the model, with input images on the left and corresponding parts on the right. Visually, the learned parts are sparse and rather complete to facilitate object recognition in the higher visual layer. We experimented with a doll image (177×314 pixels), the extracted parts (see Fig. 4.2(a)) include a left arm, a right arm, a head, a left leg, a right leg and some nonsense parts. Here, the RFs size of V4 M-cells is calculated as 71×71 pixels in accord with the image size. Giving another doll image (248×327 pixels) to the model, the model identified the same corresponding parts (see Fig. 4.2(b)). Generalizing input images to include other kinds of objects, the model learned a head, a body, two front wings and two back wings as parts of a plane image (352×447 pixels) and a fighter image (263×358 pixels), which shows that our model is stable and generalizes well (see Fig. 4.2(c)-(d)). The parts extracted for the lizard image (282×328 pixels) are two front feet, two back feet, a head and a tail (see Fig. 4.2(e)), while the detail of toe in foot patterns are not so clear due to the blur caused by convolving the input image with V1 M-cell receptive fields.



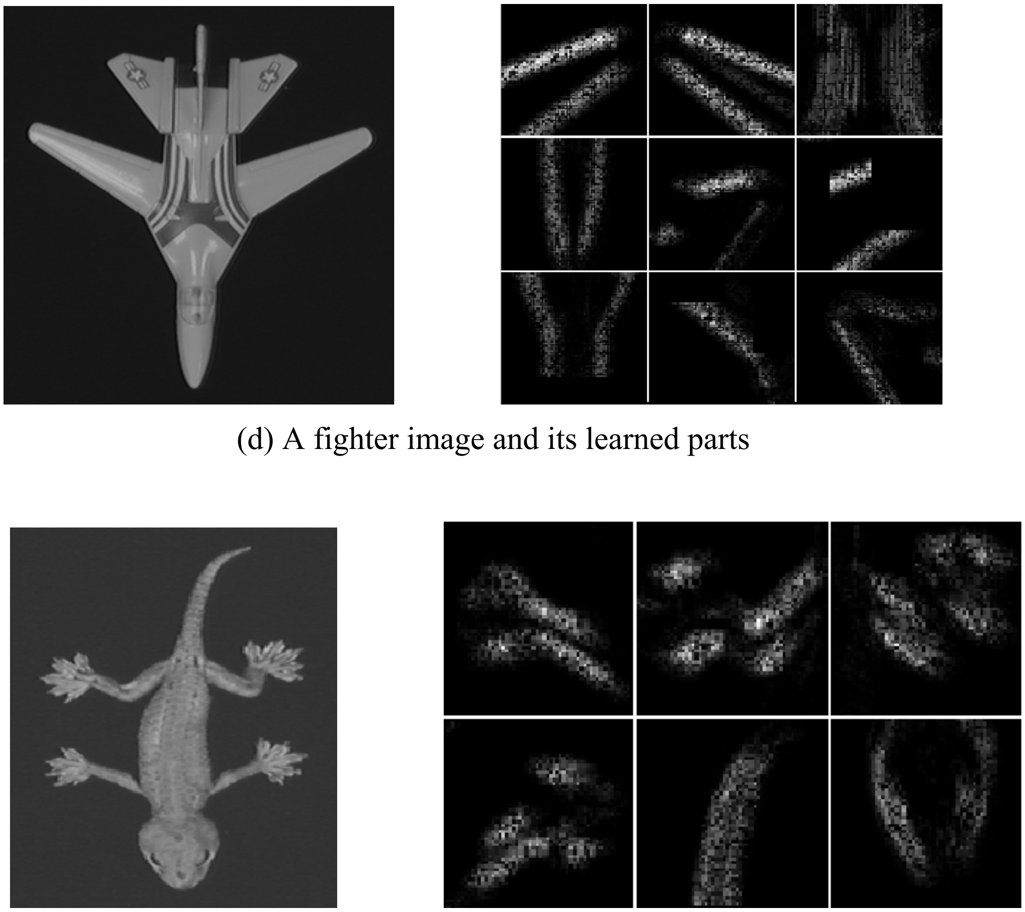
(a) A doll image and its learned parts



(b) Another doll image and its learned parts



(c) A plane image and its learned parts



(d) A fighter image and its learned parts

(e) A lizard image and its learned parts

Figure 4.2 Some examples of part decomposition by the hierarchical network. Visually, the learned parts are very sparse and rather complete**.**

4.3.4 Discussion

Anatomically, in all the visual areas, as in the entire neocortex in general, information is processed by the same few kinds of cells, arranged in the same laminar/columnar structure. The uniformity of the cortex is not limited to its anatomy: functional studies reveal a limited repertoire of computational mechanisms, of which tuned RFs are probably the most ubiquitous. In neurophysiology, the RF of a cell is defined as the part of the visual field in which a stimulus must appear to elicit a response from the cell. Together with the specification of the preferred stimulus of the cell, this constitutes a useful first approximation of its input-related function. For a complete characterization of the cell's function, its context sensitivity (induced by lateral and descending links) and its projective field should also be specified. The characteristics of the receptive fields of cortical cells and their interconnection patterns (such as the topological projections between cortical areas) constrain the kind of information processing that can be supported by the cortex

Adopting the general principles of visual information processing in the ventral pathway, we presented here a hierarchical network architecture to extract object parts by sparse and unsupervised learning. The experimental results demonstrate that our model can achieve rich and robust shape representation.

Previous work has shown how to extract object parts by non-neural mechanisms(Burl, Weber et al. 1998; Walker, Cootes et al. 1999; Nelson 2003) and by neural network mechanisms (Becker and Plumbley 1996; Lee and Seung 1999; Fiser and Aslin 2001; Li, Hou et al. 2001; Edelman, Hiles et al. 2002; Edelman, Intrator et al. 2002). Our work is different from related models by the integration of sparse representation and unsupervised learning properties in the hierarchical network. Hierarchical organization is beneficial because different modules need not duplicate the machinery required for the shared components of the analysis. Moreover, unsupervised learning allows the visual system to automatically extract parts for novel objects without employing any prior knowledge. Sparse representation has intrinsic fault-tolerance and low-power consumption potential, which are crucial properties for implementing any physical system.

We have demonstrated that our model automatically decomposes objects into parts and successfully achieves stable, efficient and natural description of shapes. However, objects with different complexity may need different amounts of parts to represent them. Especially considering the importance of using resources efficiently in the competition for survival, it is reasonable to think that the model should adaptively extract the least and most informative parts to reliably represent the object shape. In other words, the improvement for inhibiting nonsense and less meaning parts is very desirable in the model.

Our experimental results also indicate that the model learned and tested is not rotation invariant, because rotation causes significant differences between the object's actual shape and its projected images. However, we typically see objects from many different perspectives and manage to recognize them well despite the variations in appearance, though not so easily with arbitrary rotation. Our intuitions about such shape constancy are based primarily on situations in which we perceive the same object from different station points by continuously moving from one view to another. Under these conditions, the perspective changes, from moment to moment, are quite gradual and easily perceived as shape preserving, especially when the object is being viewed continuously. Moreover, the image sequence caught by a camera almost varies smoothly over time. To take advantage of these general properties of the physical world, our challenge for the future is developing viewpoint invariant representations of object shapes by capturing the temporal relationships among the input image sequences.

4.4 Comparison of V4 Model and Biological Cells

Understanding how the neural population represents shape information and how such representations arise within the cortex is one of the main objectives of visual neuroscience. The ventral pathway in the visual cortex is responsible for the perception of shapes. Area V4 is an important intermediate stage of the visual pathway and provides the major input to the final stages in the IT for object recognition**.** A number of physiological studies have shed some light on the ventral stream in primate visual cortex (Desimone, Schein et al. 1985; Gallant, Braun et al. 1993; Kobatake and Tanaka 1994; Schiller 1995; Gallant, Connor et al. 1996; Subirana-Vilanova and Richards 1996; Wilkinson, James et al. 2000; Larsson, Amunts et al. 2002), but most computational issues have yet to be resolved. Cells in the early stages of the ventral pathway have small RFs and are selective to simple features, such as edge orientation, while cells far along the pathway in inferotemporal cortex (IT) have large receptive fields and are selective to complex objects like faces and hands (Tanaka, Saito et al. 1991; Logothetis, Pauls et al. 1995; Kobatake, Wang et al. 1998). The general selectivity at these two stages of the ventral pathway is relatively well understood. However, cells at intermediate stages, between V1 and IT, have not been fully characterized.

One recent study of V4 (Pasupathy and Connor 2001) investigated the nature of shape representation in V4 cells of the primate ventral pathway. Building on a previous study (Pasupathy and Connor 1999) in which they found tuning in V4 cells to angle orientations, Pasupathy and Connor examined shape representation of V4 cells using a set of simple closed shapes formed by combining convex and concave boundary elements. They characterized a subpopulation of V4 cells as having selectivity for object-centered, position-specific, boundary conformation, such as cells that were tuned to multiple curvatures at specific angular positions from the object's center of mass.

Here, we extend our part decomposition model developed in Section 4.3 by integrating position information to equip V4 M-cells with "what+where" receptive fields, and compare the tuning properties of M-cells to those physiological measurements of V4 biological cells by a "MAX" pooling mechanism (Riesenhuber and Poggio 1999). For the "MAX" mechanism, cell response would be determined by the most active afferent and, hence, would signal the best match of any part of the stimulus to the afferents' preferred feature. Our results show that the computational model is able to reproduce the reported physiological findings (Pasupathy and Connor 2001). Similar to V4 cells, the M-cells respond to a wide variety of shapes but do not appear to encode any single type of a global shape. The shapes evoking strongest responses are characterized by a consistent type of boundary elements at a specific position within the stimulus.

4.4.1 Integration of "What" and "Where" Information

The visual system is frequently conceptualized as consisting of two functionally distinct systems: a ventral stream from area V1 to area IT and a dorsal stream from area V1 to the parietal cortex (Felleman and Van Essen 1991). Initially, the two systems were differentiated in terms of the attributes that they encode, with the dorsal stream involved in the processing of "where" information--stimulus location and motion and the ventral stream involved in the computation of "what" information--object attributes, such as shape, texture, color etc (Desimone, Schein et al. 1985; Tanaka, Saito et al. 1991). Recently, however, (Sereno and Maunsell 1998) found that cells of the lateral intraparietal area can show shape selectivity, even in a passive fixation task that does not require a motor response. (Beeck and Vogels 2000) showed that IT cells can code for the position of stimuli in the central region of the visual field and the position sensitivity of IT cells displays invariance to shape and size changes. Meanwhile, single-cell studies (Moran and Desimone 1985; Desimone and Duncan 1995; Luck, Chelazzi et al. 1997; Reynolds, Chelazzi et al. 1999; Reynolds, Pasternak et al. 2000; Reynolds and Desimone 2003) have shown that attention is capable of modulating the responses of cells at a spatial scale that is smaller than the scale of an individual RF, especially the spatial attention can modulate V4 responses when only one stimulus is locate inside the RF. Overall, these findings indicate that there is integration of “what” and “where” information in the higher visual areas. This raises the question of how to integrate “what” information with “where” information.

Many models have proposed that lateral connections are crucial for developing spatial or position tuning (Cannon and Fullenkamp 1991; Polat and Sagi 1993; Polat and Sagi 1994). Physiologically, long-range lateral connections form a dense network within the cortex. Each connection extends over several millimeters and gives rise to clusters of axon endings at regular intervals. The higher the area, the longer the range of lateral connections (Gilbert and Wiesel 1979; Schwark and Jones 1989). Lateral connections are also found to be remarkably ordered: they primarily connect areas with similar properties, such as cells with the same orientation or eye preference in the visual cortex (Gilbert, Hirsch et al. 1990; Löwel and Singer 1992). Such organization is neither genetically determined, nor static, but develops cooperatively and simultaneously with the thalamocortical afferents, and changes dynamically according to visual experience throughout life (Löwel and Singer 1992; Burkhalter, Bernardo et al. 1993). Given these findings, we believe that lateral connections are relevant to the understanding of cell response properties and plasticity higher up in the ventral stream, and therefore may play a role in integrating the “what” and “where” information in areas V4 and IT.

4.4.2 Methods

Considering that cells in V1 and V2 are not remarkably different in most respects, we did not include the V2 layer in the previous hierarchical network (see Fig. 4.1) to minimize the computation time. In the V1 layer, the training set consists of 16×16 pixel patches randomly sampled from twelve monochrome natural images. Using the learning rule in

Eq. (3.26), we pre-learned 40 complex M-cells/subspaces with the subspace dimension of 4 (ref. Section 3.3.1). The learned basis vectors associated with a single complex M-cell have approximately the same orientation and frequency.

Taking the output of the V1 layer as input, the synaptic weights onto V4 M-cells are first trained by the unsupervised learning algorithm specified by Eq. (4.3~4.5), which is distinguished from other methods by its use of non-negative constraints. These constraints lead to the sparseness of basis images (the receptive fields of V4 M-cells), because they allow only additive, not subtractive, combinations, therefore reducing the redundancy among the data (Lee and Seung 1999). Spatial ("where") information is integrated into the model by lateral connections among the V4 "what+where" M-cells as:

$$a_i = a_i - \beta \sum_{j \neq i} v_{ij} a_j \qquad (4.7)$$

and

$$v_{ij} = G(d_{ij}, K\delta) - G(d_{ij}, \delta)$$

where $1 \leq i, j \leq N$ and N is the number of V4 M-cells; a_i is the basis image (or the weight of a i-th V4 model cells) ; w_{ij} is the strength of the lateral connection between model cells i and j ; $G(d, \delta)$ is the value at d of a Gaussian with δ width. Such lateral connectivity creates a window corresponding to the location of a V4 M-cell, so as to endow the "what+where" RF to a M-cell. V4 M-cells are arranged over the afferent image by shifting the center of the Gaussian.

With the objective function converging to a small value, V4 M-cells are tuning to some boundary-based parts. Each learned part is characterized by a boundary element and position, hence, could be considered a point in a two dimensional space ($b \times p$). Each stimulus could be considered a collection of such m points/parts. A nonlinear maximum operation ("MAX") (Riesenhuber and Poggio 1999), where the strongest afferent determines the postsynaptic response, seems compatible with neurophysiological data (Riesenhuber and Poggio 1999) and is used to pool the output of V4 M-cells:

$$r = \max_{m} \left[k \cdot e^{-y_{ib}^2 / 2\delta_{ib}^2} \cdot e^{-y_{ip}^2 / 2\delta_{ip}^2} \right] \qquad (4.8)$$

where k represents the amplitude of the template matching Gaussian transfer function; y_{ib} represents the shape/what similarity measure between the i-th learned part and the tuning part of a V4 M-cell; and y_{ip} represents the position/where similarity measure between them. Hence, the response would signal the best match of any part of the stimulus to the afferent's preferred feature.

4.4.3 Experiments and Results

The stimulus set used for the comparison experiments based on the physiological studies of V4 cells was provided by Dr. Pasupathy and is shown in Fig. 4.3(a). The stimuli were constructed by systematically combining convex and concave boundary elements to form closed shapes. These boundary elements included sharp convex angles, medium and broad convex curves, and medium and broad concave curves. Each shape was defined by the number and configuration of convex projections it contained. Corresponding to Pasupathy and Connor's physiological studies, the stimulus that elicited the highest

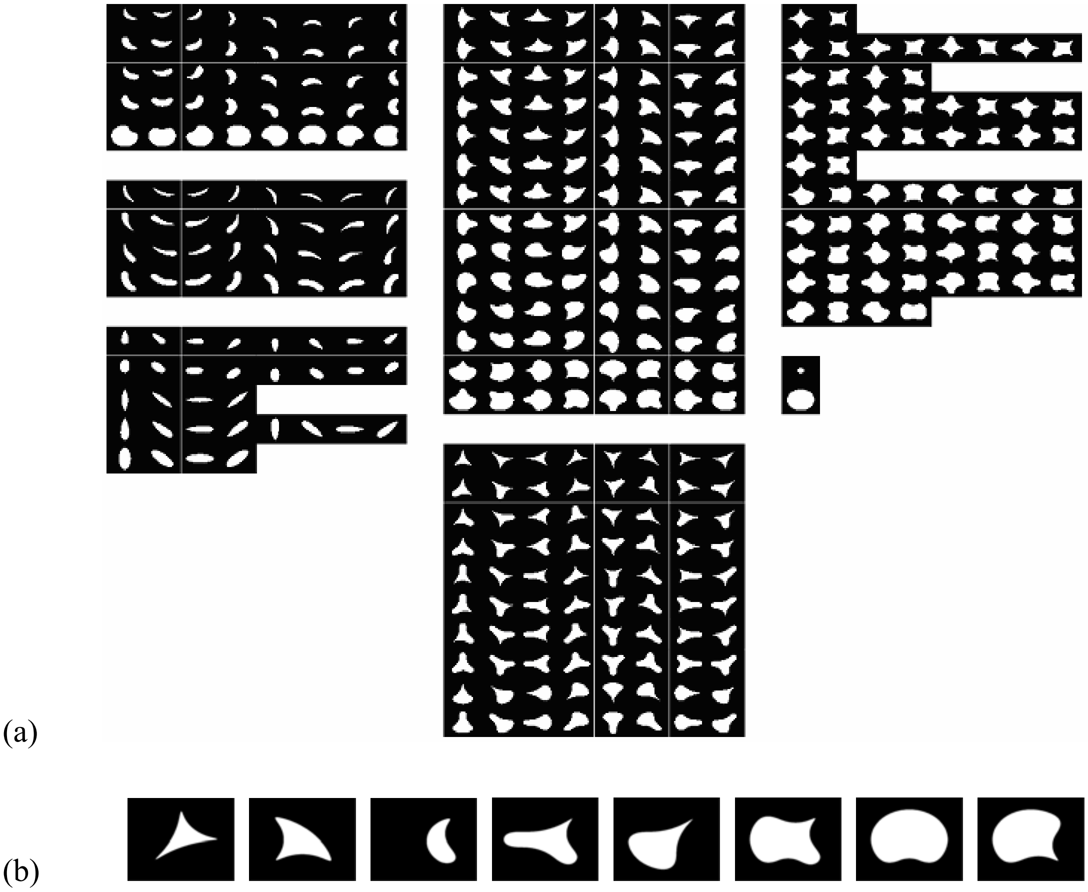


Figure 4.3 (a) The stimulus set used for computational simulations. Each stimulus is represented by a white icon drawn within a black rectangle. (b) The effective stimuli that elicited the highest response in Pausupathy and Connor's physiological experiments were selected and rotated to train the tuner parts of V4 M-cells.

response of each given V4 cell was selected as the M-cell's "tuner" in Fig. 4.3(b) and is rotated $\pm 5^\circ$, $\pm 10^\circ$ and $\pm 15^\circ$ about the z-axis to bring dynamics into the model, which constitute the effective stimuli for training the tuning part of each experimental V4 M-cell.

In the V4 layer, the training data $X = \{CC_j, j = 1,...,40\}$ is obtained by using Eq. (3.29) to convolve the stimulus image (512×422 pixels) with pre-learned V1 complex M-cell RFs. In line with the known visual topography (Wallis and Rolls 1997): RF sizes of cells in the ventral pathway become larger by a factor of approximately 2.5 in a succeeding stage and the RF sizes of cells in IT area can include the entire central visual field, we choose the RF size for a V4 M-cell as $RF_4\_SIZE = \sqrt{\frac{l_I^{\,2}}{2.5^2}}$ with $l_I = 422$, the small dimension size of a stimulus image (width×length). Starting with random initial conditions for A and S, the unsupervised learning algorithm had mostly converged after 30 iterations. This process took 5 minutes running Matlab on a Dell Precision Station (530MT, 2GHZ and 4GB). For comparing the tuning properties of V4 M-cells with cells in physiological experiments (Pasupathy and Connor 2001), we use Eq. (4.8) to characterize the output of M-cells. In Eq. (4.8), the standard deviation (SD) of shape/what similarity is chosen to be 0.42 based on physiological data (Pasupathy and Connor 2001), implying responsiveness to a range of boundary elements. Accordingly, the SD of position/where similarity is chosen to be 34°, implying sensitivity to boundary elements at fairly restricted location.

Figure 4.4 shows the responses to the stimulus set of a V4 cell and a V4 M-cell. Both responses are linearly scaled between 0 and 1. The darker the background the higher the response exhibited to that stimulus. The V4 cell response is adapted from Figure 2 of (Pasupathy and Connor 2001). Visually, the tuning properties of the given cell and this experimental M-cell are very similar. To quantitatively compare the learned and actual response[3], the M-cell response is plotted against the V4 cell response in the lower right

[3] Dr. Anitha Pasupathy suggested that they are still working with the recording data of V4 biological cells and aren't ready to part with them yet. Therefore, we used Photoshop to acquire the gray level surrounding each stimulus icon in Pasupathy's figures. Here, the gray level denotes average cell response exhibited to the stimulus (Pasupathy and Connor 2001).

portion of Fig. 4.4(b). The scatter plots represent goodness-of-correspondence for tuning properties of the M-cell and the V4 cell. The vertical axis represents the actual response of the V4 cell and the horizontal axis represents the response of the M-cell. The correlation coefficient of 0.81 between learned M-cell responses and actual V4 cell responses suggests a strong correspondence between their tuning properties.

For both cases, stimuli that evoked strong responses varied widely in overall structure and included crescents, triangles, teardrops, and four-pronged shapes. A common feature of these shapes, however, was the presence of a convex projection near the bottom left (relative to the objective center). Stimuli with a sharp convex curve at this position were particular effective, while stimuli with a medium convex curve evoked moderate responses. Thus, the boundary element in a specific region of the object (bottom left and bottom) is clearly a major determination of this M-cell and its corresponding V4 cell responses to complex shapes.

Figure 4.5 shows another comparison between a V4 biological cell and a M-cell. This second V4 cell and its corresponding M-cell both are sensitive to boundary elements on the right side of the object, responding best to concave curvature at that position. The correspondence between the V4 cell response and the M-cell response is represented by the scatter plots in lower right of Fig. 4.5(b) and reflected by the correlation coefficients of 0.78.

Figure 4.6 shows the third comparison. The third V4 cell and its corresponding M-cell both are sensitive to boundary elements at the top right, responding best to sharp convexity. The correlation coefficient between the V4 cell response and the M-cell response is 0.73.

The fourth V4 cell and its corresponding M-cell shown in Fig. 4.7 are tuned to a broad convex at the top of the object. The correlation coefficient between this V4 cell response and the M-cell response is 0.71. These results show that our computational model is able to reproduce Pausupathy and Connor’s physiological findings: V4 cells responded to a wide variety of shapes but did not appear to encode any single type of global shapes. The shapes evoking strongest responses were characterized by a consistent type of boundary elements at a specific position within the stimulus.

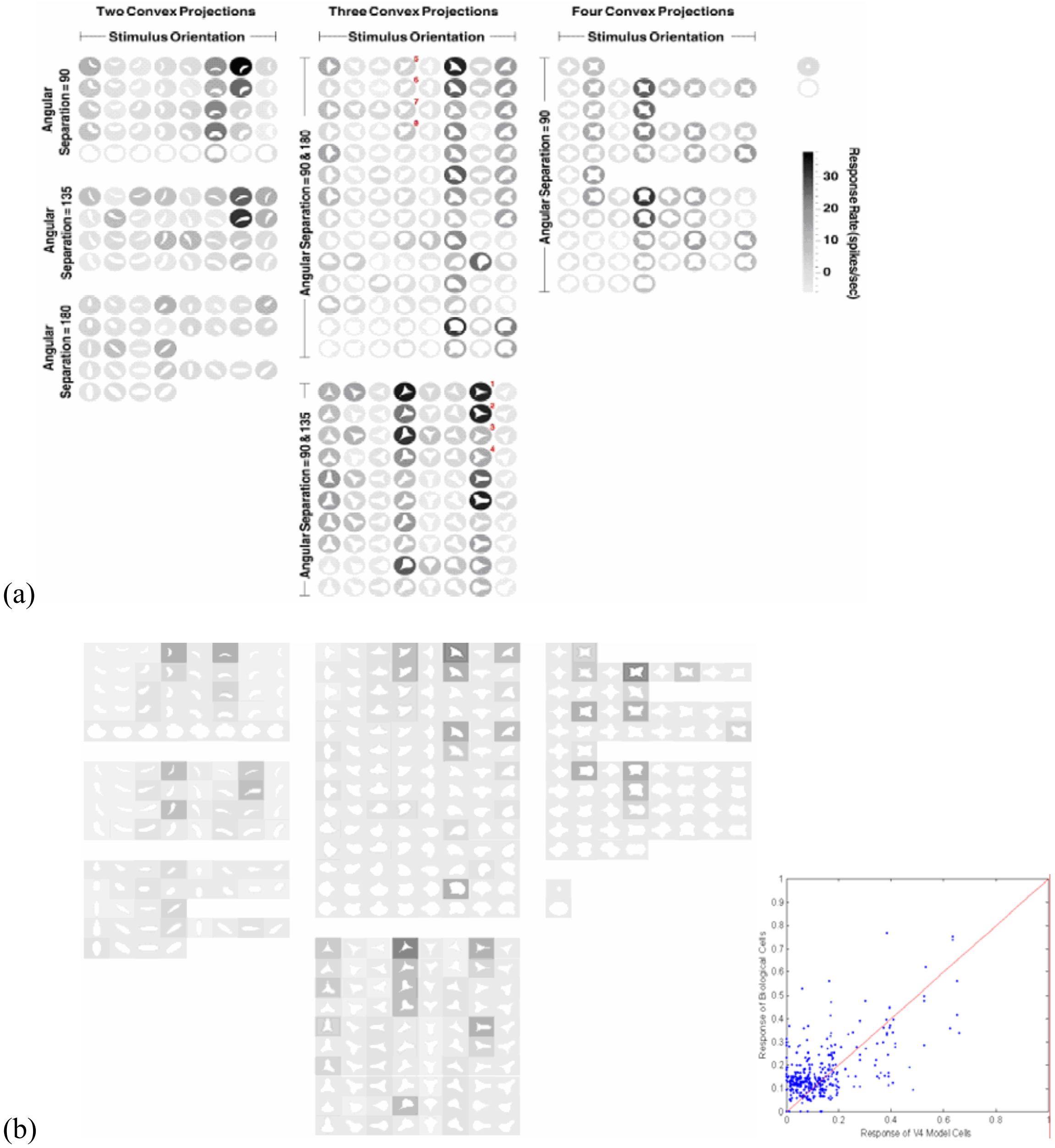


Figure 4.4 V4 biological cell and M-cell response to the stimulus set. (a) V4 cell response adapted from Figure 2 of (Pasupathy and Connor 2001). (b) The response of a M-cell to each stimulus is shown in terms of the shading surrounding the stimulus and has been linearly scaled between 0 and 1. The darker the background the higher the response exhibited to that stimulus. The M-cell response is plotted against the V4 cell response in the lower right.

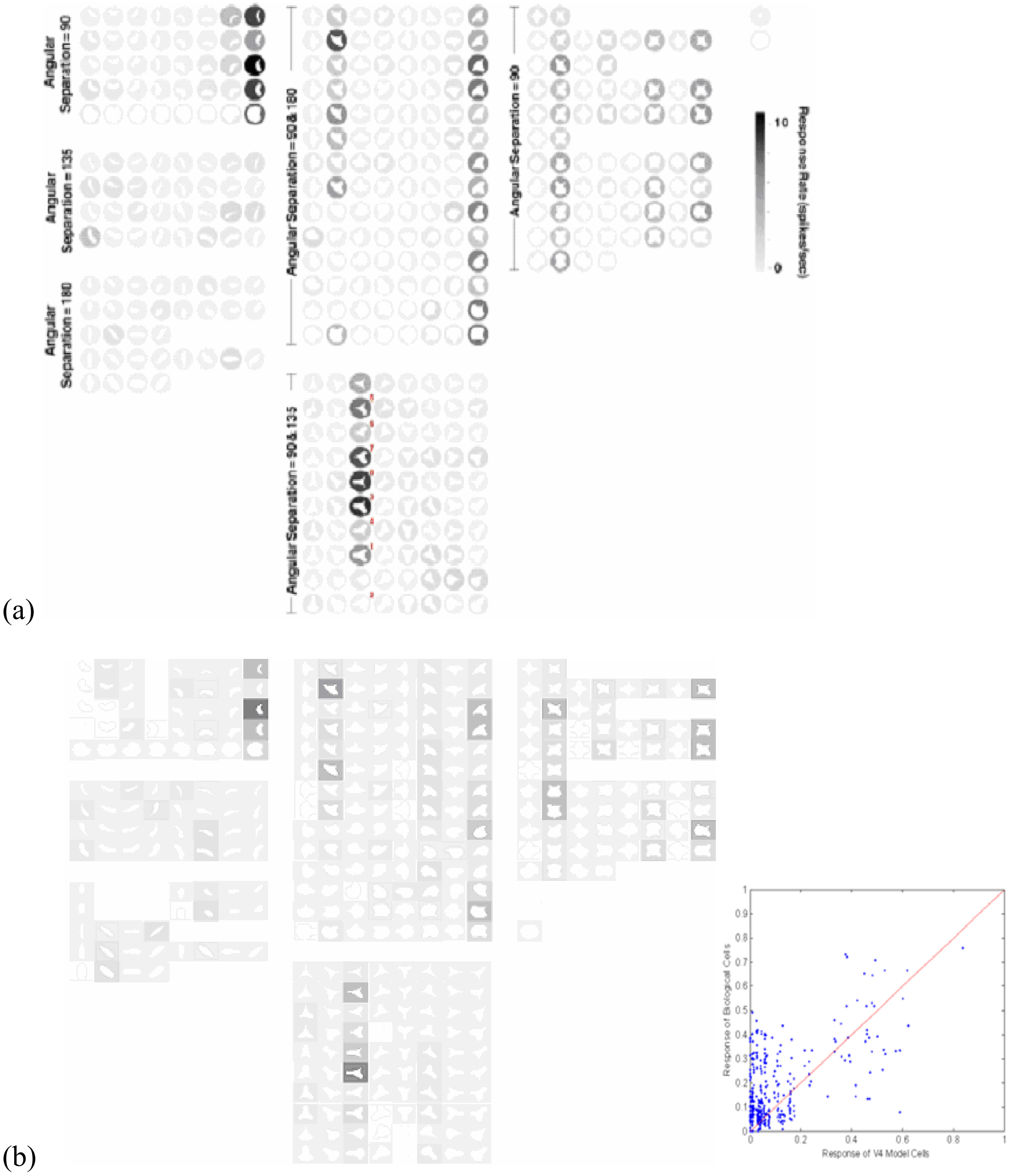


Figure 4.5 This V4 cell and its corresponding M-cell both are sensitive to boundary elements on the right side of the object, responding best to concave curvature at that position. (a) V4 cell response adapted from Figure 4 of (Pasupathy and Connor 2001). (b) The response of a M-cell. Scatter plots in lower right showing correspondence between the responses of the V4 cell and M-cell.

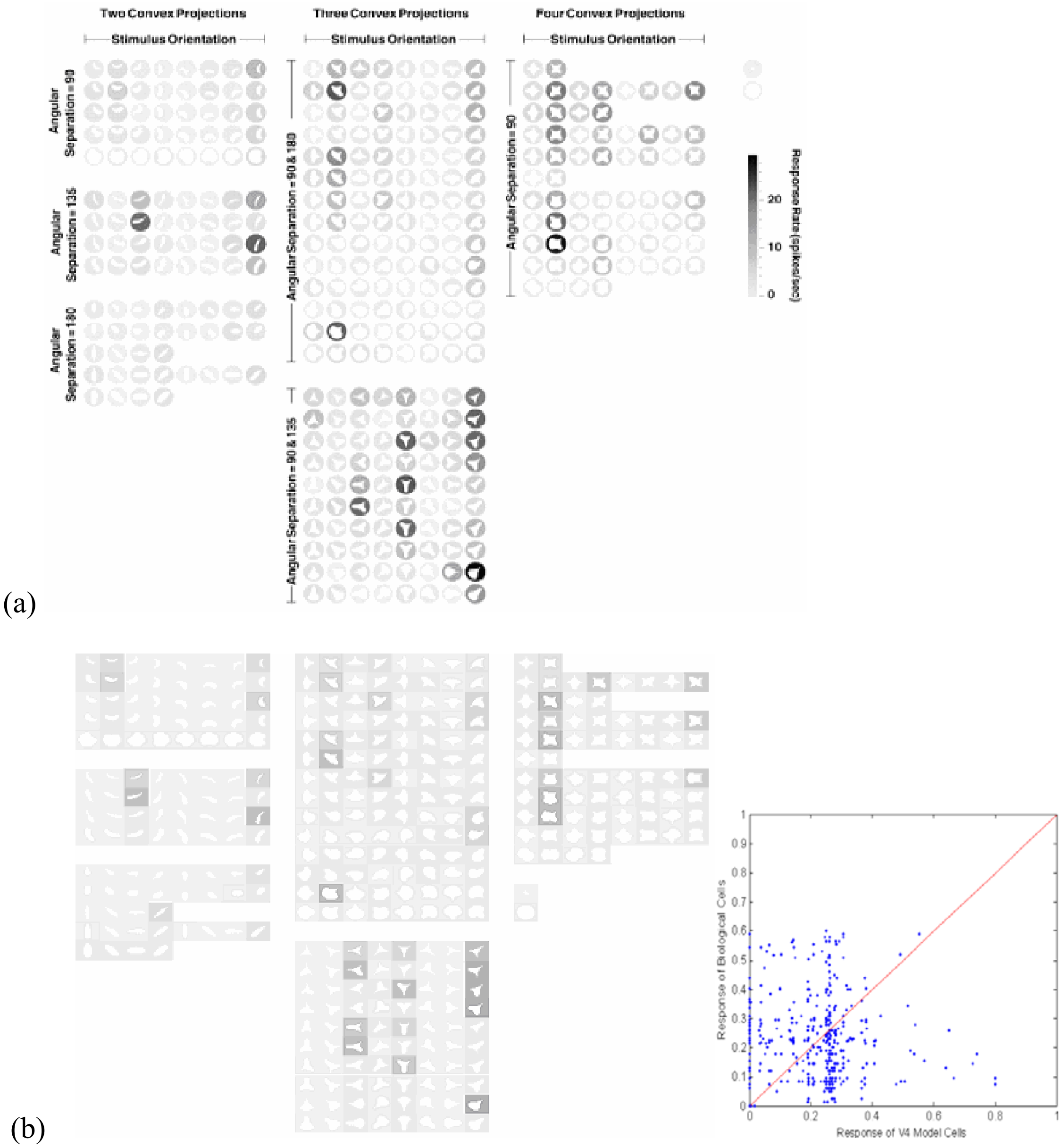


Figure 4.6 The V4 cell and its corresponding M-cell both are sensitive to boundary elements at the top right, responding best to sharp convexity. (a) V4 cell response adapted from Figure 5 of (Pasupathy and Connor 2001). (b) The response of a M-cell. Scatter plots in lower right showing correspondence between the responses of the V4 cell and M-cell.

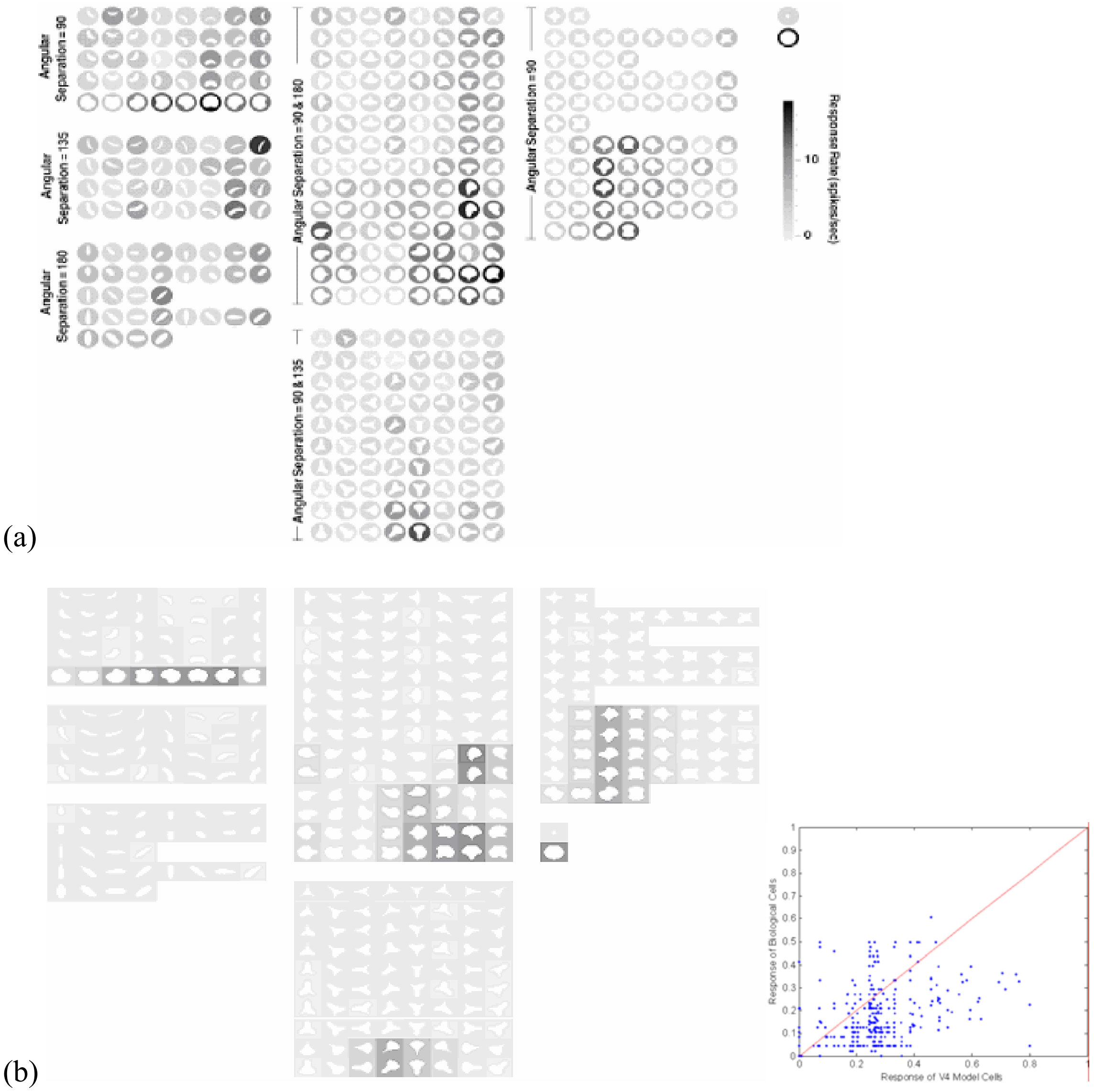


Figure 4.7 The V4 cell and its corresponding M-cell both are tuned to broad convex at the top. (a) V4 B-cell response adapted from Figure 8 of (Pasupathy and Connor 2001). (b) The response of a V4 M-cell. Scatter plots in lower right showing correspondence between the responses of the V4 B-cell and M-cell.

To verify that responses did not depend on some specific placement of stimuli relative to the RF, we tested shape tuning of V4 M-cells at multiple positions. We selected two stimuli based on Pasupathy's studies for directly comparing position tuning of the Fig. 4.6 M-cell and V4 biological cell. The effective star-shaped stimulus contains the boundary element (i.e. a sharp convexity at the top right) that drove the cell and another ineffective stimulus does not contain this boundary element. We presented each stimulus at 25 positions evenly arranged in a 5×5 grid centered on the RF. The test results are shown in Fig. 4.8. The effective stimulus evoked strong responses at multiple positions, while the ineffective stimulus never evoked a strong response. Same as the biological measurements by Pasupathy and Connor, the stimulus containing the critical boundary element evoked the strongest response across positions, demonstrating that the M-cell is more than merely template-based tuning, but showing position invariance to the preferred boundary element.

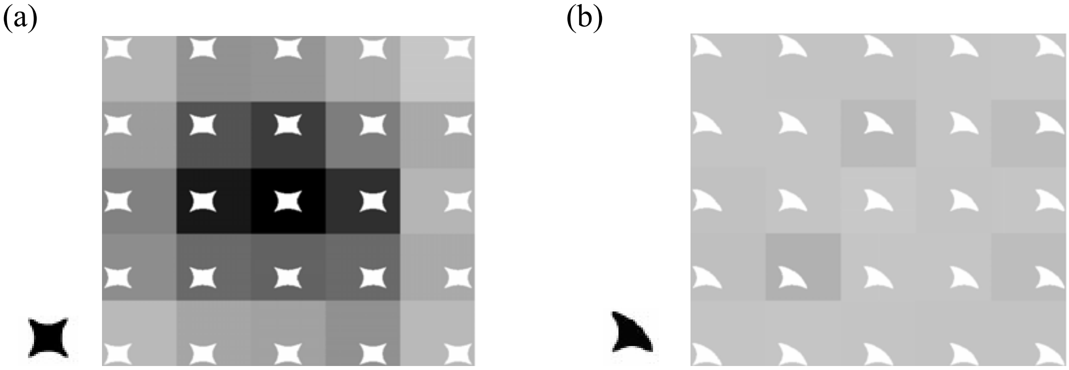


Figure 4.8 Position test for the Fig. 4.6 M-cell. Surrounding gray levels denote average responses to an effective (left) and ineffective (right) stimulus, both shown in black at left, presented at 25 positions on a 5×5 grid centered on the RF.

4.4.4 Discussion

We presented a hierarchical network architecture inspired by the visual ventral pathway to extract object parts using sparse and unsupervised learning. A hierarchical organization is beneficial because different modules need not duplicate the machinery required for shared components of the analysis. Unsupervised learning allows the visual system to automatically extract parts of novel objects based on statistical properties, and without employing any explicit knowledge. Sparse representation has intrinsic fault-tolerance and low-power consumption potential, which are crucial properties for implementing any physical system. Moreover, our previous results have demonstrated that this part decomposition model can automatically decompose objects into useful parts and successfully achieve stable, efficient and natural representations of shapes.

In this section, we extended the part decomposition model by integrating with position (”where” information) consistent with the belief that such “where” information is crucial for object recognition. Comparing the tuning properties of model cells to those of V4 biological cells by “MAX” pooling mechanism (Riesenhuber and Poggio 1999), we found that the response properties of M-cells are similar with those of V4 cells reported in Pasupathy and Connor’s psychological experiments.

Our model represents a biologically plausible mechanism that exhibits the same object-centered, position-specific curvature tuning found by Pasupathy and Connor in a subpopulation of real V4 cells. Consequently, it may provide inspiration to further explanation of shape representation in biological settings.

However, the response properties of model cells still show some variation from those of reported V4 cells. One possible reason is that the optimal standard deviations in Eq. (4.8) are not determined based on minimizing the squared error between the responses of M-cells and V4 biological cells in physiological experiments. Another possible reason is that the tuned parts of M-cells are learned from limited training data. The RF eccentricities of V4 M-cells in our model are assumed to be near the fovea, but those of V4 cells are not, which is another possible source of variation.

4.5 Validating the reliability of Learned Parts via SOM

We have demonstrated that our model can automatically decompose objects into parts and successfully achieve stable, efficient and natural representation of shapes. A major problem of the model is that the reliability of the learned parts is not known. The model gives a specified number of parts, which include both key parts and non-sense (noise) parts, but it is not known which are to be taken seriously, or which are key parts for object recognition in higher visual area.

As with any statistical method, it is necessary to analyze the statistical reliability (significance) of the obtained components. The finite sample size induces statistical errors in the estimation, this is a situation where classical analysis of statistical reliability should be used (Himberg, Hyvärinen et al. 2004). Another problem of our model is that the learning algorithms in V1 and V4 layers have random (stochastic) elements, i.e. the algorithms give somewhat different results for every run of the algorithm. All these statistical algorithms are based on the minimization or maximization of an objective function, such as likelihood, mutual information, or negentropy. The randomness of the algorithm stems from the fact that it is not always possible to find the point that globally minimizes the objective function, because the objective function used in the estimation may have many local minima. Consequently, most such algorithms may not always perform properly, such as getting stuck in local minima with strongly suboptimal values of the objective function. Also, the basic principle of the stochastic algorithms related to gradient descent is to start in some initial point, and then make steps in a direction that decrease the objective function, until one finds a point in which the objective function has been minimized. Depending on the point where the search was started (the "initial point"), the algorithm will find different local minima. In the case of a very high-dimensional signal space, the probability of finding the global minimum may be very small. Thus, it is reasonable to investigate the reliability of the parts learned by our model.

Here, we use the temporal SOM algorithm to project the parts learned on a continuous time scale into the signal space, based on the similarity or correlation among them. Each learned part is a point in the signal space. The parts are grouped into a number of disjoint clusters, which represent the same or similar parts should be close to each other within the same cluster. The more similar the parts, the closer the points are.

4.5.1 Self Organizing Map (SOM)

The Self Organizing Map (SOM) is one of the best known unsupervised neural learning algorithms (Kohonen 1995), which has been used successfully to analyze very large files in various fields, such as process monitoring and visualization (Simula and Kangas 1995), exploratory data analysis (Ultsch and Siemon 1990) and simulation of brain-like feature maps (Kohonen and Hari 1999). SOM allows easy visualization of complex data and is robust to minor experimental variation. The goal of the SOM algorithm is to find prototype vectors that represent the input data set and at the same time realize a continuous mapping from the input space to a two-dimensional space.

Each cell i in the SOM is associated with a weight vector $w_i = [w_{i1}, w_{i2}, \cdots, w_{in}]^T$ in the map space $V_M \in R^2$, which has the same dimension as the stochastic input vector

$X = [x_1, x_2, \cdots, x_n]^T$ from input space $V_I \in R^m$. Through an unsupervised learning process, the output cells are tuned and organized after several presentations of the data. The learning algorithm that leads to self-organization can be summed in two steps:

(i) A winning or best-matching unit, denoted by bmu, of the map is found by using a given similarity measure (Euclidean distance, for example) between the input and weight vectors:

$$bmu = \arg\min_{i \in V_M} \{\|x(n) - w_i(n)\|\} \tag{4.9}$$

or, equivalently, $\|x(n) - w_{bmu}(n)\| = \min_{i \in V_M} \{\|x(n) - w_i(n)\|\}$ (4.10)

(ii) The winner and its neighbors in the map have their weights $w_i(n)$ updated towards the current input $x(n)$ according to

$$w_i(n+1) = w_i(n) + \gamma(n) h_{bmu,i}(n)(x(n) - w_i(n)) \tag{4.11}$$

where $i \in W_M$ and $\gamma(n)$, $0 \le \gamma(n) \le 1$, is a scalar valued adaptation gain. We assume that updating of the weights is done incrementally after each input vector $x(n)$ presentation. The initial values of $w_i(0)$ are usually random. The neighborhood function, $h_{bmu,i}(n)$, is a smoothing kernel defined over the cells in the array, and plays a central role in the learning process. A typical choice for $h_{bmu,i}(n)$ is a Gaussian function $h_{bmu,i}(n) = \exp(-\|r_i - r_b\|^2 / \sigma(n)^2)$, where σ controls the width of the function and r_i and r_{bmu} are the SOM index vectors of the unit i and the best matching unit bmu.

During learning the function $h_{bmu,i}(n)$ normally approaches a delta function, i.e., σ slowly approaches zero as training progresses. When good quantization is desired, the map should be trained with only bmu in $h_{bmu,i}(n)$ once the map has organized. During this quantization stage the gain has to be sufficiently small to avoid losing the map ordering, how small varies from case to case. In order to guarantee convergence of the algorithm, the gain, $\gamma(n)$, should decrease as a function of time or training steps according to conditions:

$$\lim_{t \to \infty} \int_0^t \gamma(t')dt' = \infty$$

and $$\lim_{t\to\infty}\int_0^t (\gamma(t'))^2 dt' = C, C < \infty \tag{4.12}$$

If the map is trained properly, i.e the gain and the neighborhood functions are properly decreased over training a mapping is formed, where weight vectors specify the cluster centers satisfying the vector quantization criterion:

$$E = \min\left\{\sum_{j=1}^{M}\left\|x_j - w_{bmu}(x_j)\right\|\right\} \tag{4.13}$$

where we seek to minimize the sum squared distance E of all input patterns, $x_j, j = 1,\cdots M$, to the respective $bmus$ with weight vectors $w_{bmu}(x_j)$.

One might say that the SOM is a nonlinear projection of the probability density function $p(X)$ of the high-dimensional input data vector X on the two-dimensional display. During learning, i.e. the pocess in which the nonlinear projection is formed, those cells that are topographically close in the array, up to a certain geometric distance, will activate each other (cooperate) to learn something from the same input X . This will result in a local relaxation or smoothing of the weight vectors of cells in this neighborhood, which, in continued learning over time, leads to global ordering or map that preserves the topology of the input samples in the sense that adjacent patterns are mapped onto adjacent regions of the map. Owing to this topology-preserving property, the SOM is able to cluster input information and spatial relationships of the data on the map. The SOM can also the employed as an intermediate method in data clustering, providing clustering via visualization. For example, k-means can cluster the weight vectors of the SOM built from the observed data. Other methods can also be applied to find more quantitative information from the generated SOM, such as summaries and interesting groupings of map units.

4.5.2 SOM in Sequence Processing

The original SOM concept is based on the matching of static signal patterns alone, and the convergence is not assured unless the topological relationships between the different input patterns are steady. The input patterns may, in addition to being related spatially, occur in a sequence.

The temporal Kohonen Map (TKM) was proposed by (Chappell and Taylor 1993) as a modification of the original SOM. This modification is not only capable of separating different input patterns but is also capable of giving context to patterns appearing in sequences. In the TKM the involvement of the earlier input vectors in each unit is represented by using a recursive difference equation which defines the current unit activity as a function of the previous activations and the current input vector. The outputs of the normal SOM are reset to zero after presenting each input pattern and selecting the bmu with the typical winner take all strategy, hence the map is sensitive only to the last input pattern. In the TKM the sharp outputs are replaced with leaky integrator outputs which, once activated, gradually lose their activity.

The outputs in the TKM approximate the behavior of natural cells, which retain an electrical potential on their membranes with decay. In the TKM this decay is modeled with the difference equation:

$$V_i(n) = dV_i(n-1) - (1/2)\|x(n) - w_i(n)\|^2 \tag{4.14}$$

where $0 < d < 1$ can be viewed as a time constant. $V_i(n)$ is the activation of the unit i at step n, $w_i(n)$ is the reference or the weight vector in the unit i and $x(n)$ is the input pattern. Now the bmu is the unit with maximum activity. Equation (4.14) has the following general form:

$$V_i(n) = -(1/2)\sum_{k=0}^{n-1} d^k \|x(n-k) - w_i(n-k)\|^2 + d^n V_i(0) \tag{4.15}$$

where the involvement of the earlier inputs is explicit. Further analysis of Eq. (4.15) shows how the optimal weight vectors in the vector quantization sense can be solved explicitly when n is assumed to be sufficiently large to render the last residual term, corresponding to initial activity, insignificant. When w_i is assumed constant, we have:

$$\frac{\partial V(n)}{\partial w} = -\sum_{k=0}^{n-1} d^k (x(n-k) - w) \tag{4.16}$$

Now, when w is optimal in the vector quantization sense (Eq. 4.13), the derivative in Eq. (4.16) is zero as this minimizes the sum in Eq. (4.15). Hence setting the left hand side of Eq. (4.16) into 0 yields:

$$0 = -\sum_{k=0}^{n-1} d^k (x(n-k) - w)$$

$$w = \frac{\sum_{k=0}^{n-1} d^k (x(n-k))}{\sum_{k=0}^{n-1} d^k} \tag{4.17}$$

This result shows how the optimal weight vectors, in the sense of vector quantization, are linear combinations of the input patterns. Since the TKM is trained with the normal SOM training rule, it attempts to minimise the normal vector quantization criterion in Eq. (4.13), which is other than the criterion suggested by Eq. (4.17).

4.5.3 Methods

In the real world, the image sequence caught by a camera generally varies smoothly over time. For validating the algorithm reliability of the V4 model, the input image sequence consisted of continuous images with approximately 20 degrees between neighboring views. The model gave 9 learned parts for each image.

We use the TKM algorithm to project part patterns leaned in continuous time scale according to their mutual similarities. A natural measure of similarity between the learned parts is the absolute value of their mutual correlation coefficients $r_{ij}, i, j = 1, \cdots K$ based on the shape and position similarity. Each point in the signal space represents a learned part. The points are partitioned into a number of disjoint clusters.

To direct the attention to those clusters that are the most compact and interesting, we introduce a (conservative) cluster quality index, I_q, that reflects the compactness and isolation of a cluster. It is computed as the difference between the average intra-cluster similarities and average extra-cluster similarities. Let us denote by C the set of indices of all the learned parts, by C_m the set of indices that belong to the *m*-th cluster, and by $|C_m|$ the size of the *m*-th cluster. We can then define the cluster quality index as

$$I_q(C_m) = \frac{1}{|C_m|^2} \sum_{i,j \in C_m} \sigma_{ij} - \frac{1}{|C_m||C_{-m}|} \sum_{i \in C_m} \sum_{j \in C_{-m}} \sigma_{ij} \tag{4.18}$$

where $\sigma_{ij} = |r_{ij}|$ and C_{-m} is the set of indices that do not belong to the m-th cluster. Eventually, $I_q(C_m)$ is equal to one for an ideal cluster and decreases as C_m is less compact and isolated.

An important parameter that needs to be selected in any clustering method is the number of clusters L. It is reasonable to start studying the clustering with the number of clusters L equal to the data dimension k and investigate the values of cluster quality index in rank order of clusters. There are also quantitative indices for suggesting the number of clusters that best fit to the data. These include four of the Dunn-like indices in (Bezdek and Pal 1998) and the R-index defined in (Levine and Domany 2001). Empirical studies on such indices, e.g., (Bandyopadhay and Maulik 2001; Maulik and Bandyopadhay 2002) often yield different results depending on the character of the data used without no clear indication of general superiority. Here, the index is defined as

$$I_R = \tfrac{1}{L} \sum_{m=1}^{L} \frac{S_m^{in}}{S_m^{ex}} \tag{4.19}$$

where

$$S_m^{in} = \frac{1}{|C_m|^2} \sum_{i,j \in C_m} d_{ij} \quad \text{and} \quad S_m^{ex} = \min_{m' \neq m} \frac{1}{|C_m||C_{m'}|} \sum_{i \in C_m} \sum_{j \in C_{m'}} d_{ij}$$

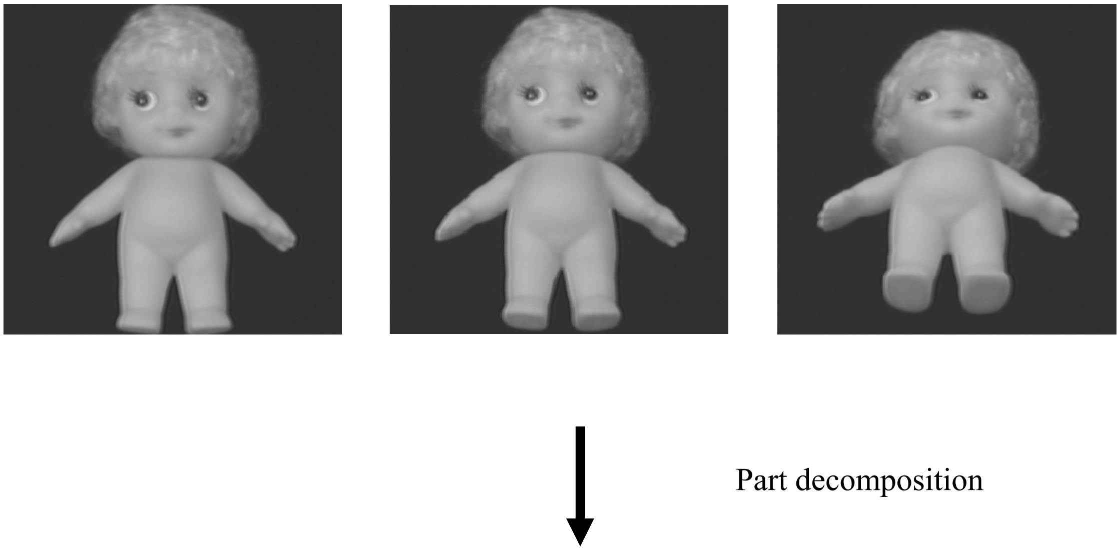
This index is a variant of the Davies-Bouldin index (Bezdek and Pal 1998). And it looks for compact and well-separated clusters, where the minimum *IR* indicates the best partition. However, we note that such an index should be used only by side of the explorative investigation.

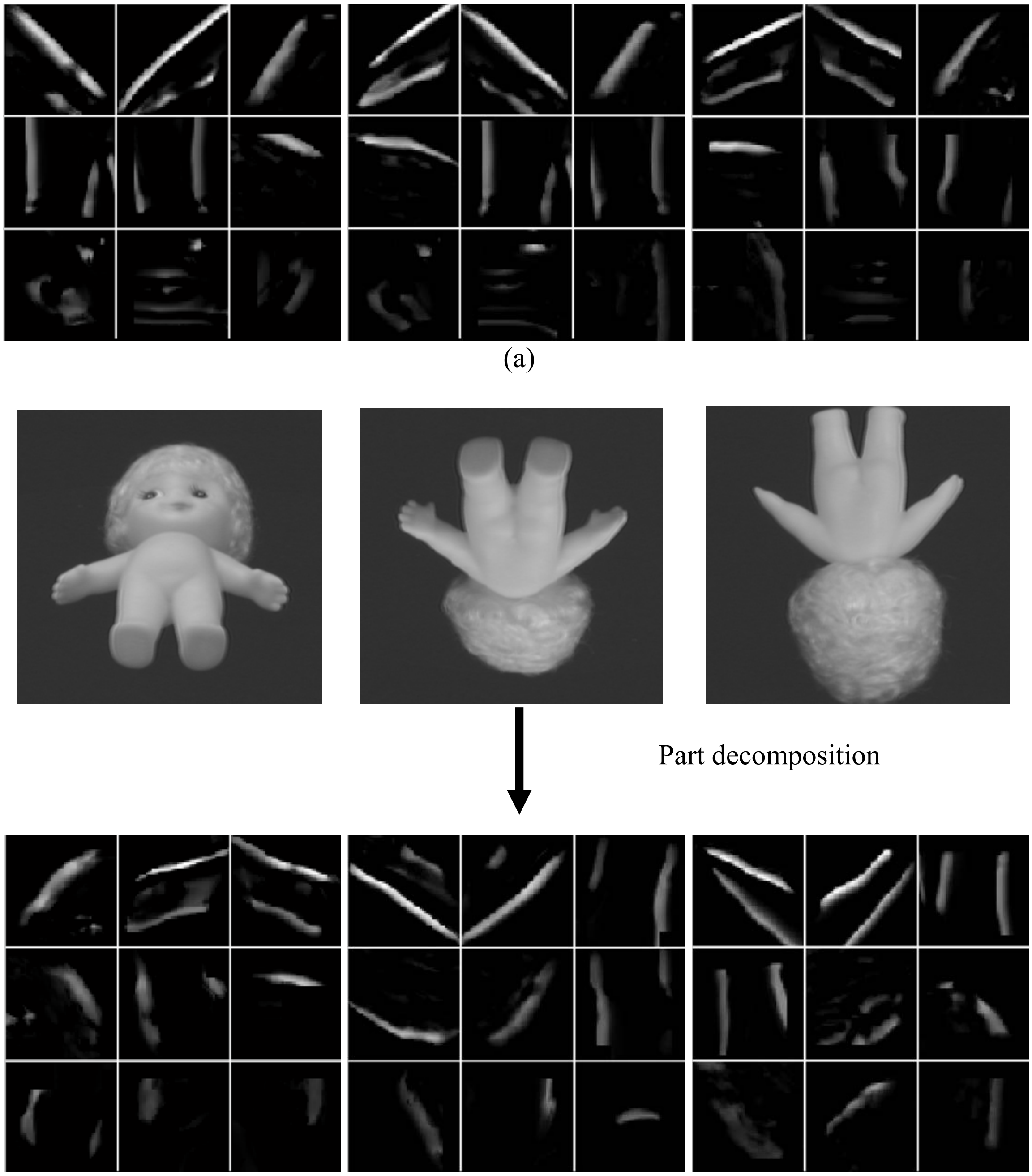
Attractively, it is possible to develop viewpoint invariance shape representation (to some extent) based on selecting “centropart” in the reliability analysis. In fact, we can integrate information over many image frames performed by computing a representative point for each tight cluster. The “centropart” is the point in the cluster that has the maximum sum of similarities (as measured by correlation coefficients) to other points in the cluster.

4.5.4 Results and Discussion

We used an the image sequence consisting of 120 rotated doll images (177×314 pixels) as input, the V4 model learned $120 \times 9 = 1080$ parts, and some examples are shown in Fig. 4.9. It can be seen that parts extracted from each image frame include at least two arms and two legs. And in fact, these are key parts that distinguish a doll from other objects. We can predict that the key parts correspond to clusters that are small and well separated from the rest of the estimates. In contrast, unreliable/noisy estimates correspond to points which are do not belong to any cluster.

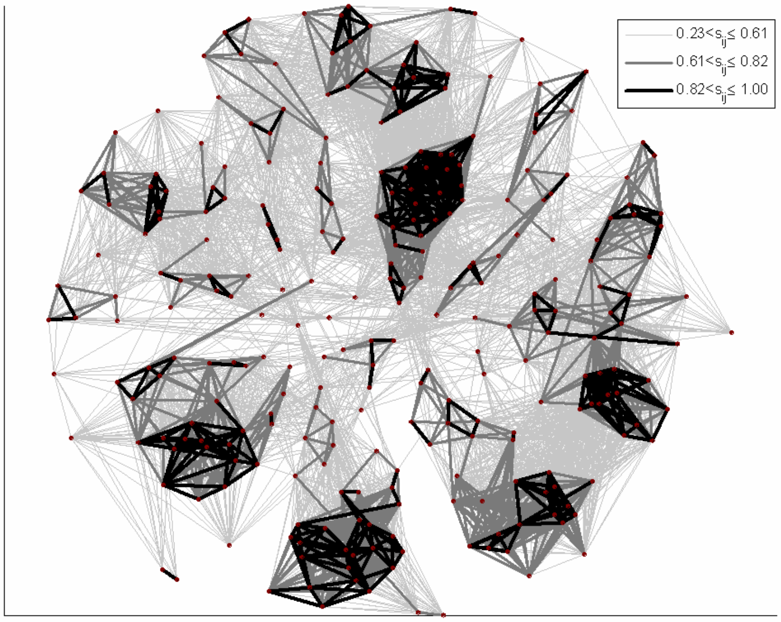
Over the first 30 input frames (over 0° ~ 45° rotations), the learned parts are projected and clustered by the temporal Kohonen Map method. We set the number of clusters to 9, which is the dimension of the learned parts for each image. The clusters and their interrelations are visualized in Fig 4.10 (a). Gray lines connect the learned parts whose similarity is larger than a threshold, the darker the line the stronger the similarity. Fig 4.10 (b) shows the quality index I_q for each cluster. The "centroparts" associated to the most robust clusters #1-5 are presented in quality rank order in Fig 4.10(c). We retest the model by taking the first 60 frames (over 0° ~ 90° rotations) as input. The validating experimental results are shown in Fig 4.11. In this case, rotation causes the slant of the doll, but the doll objects in frames still have distinguished parts, such as arms, legs and a head. As a result, the clusters corresponding to arms, legs and the top contour of a head are all ranked to the top which is quite reasonable.



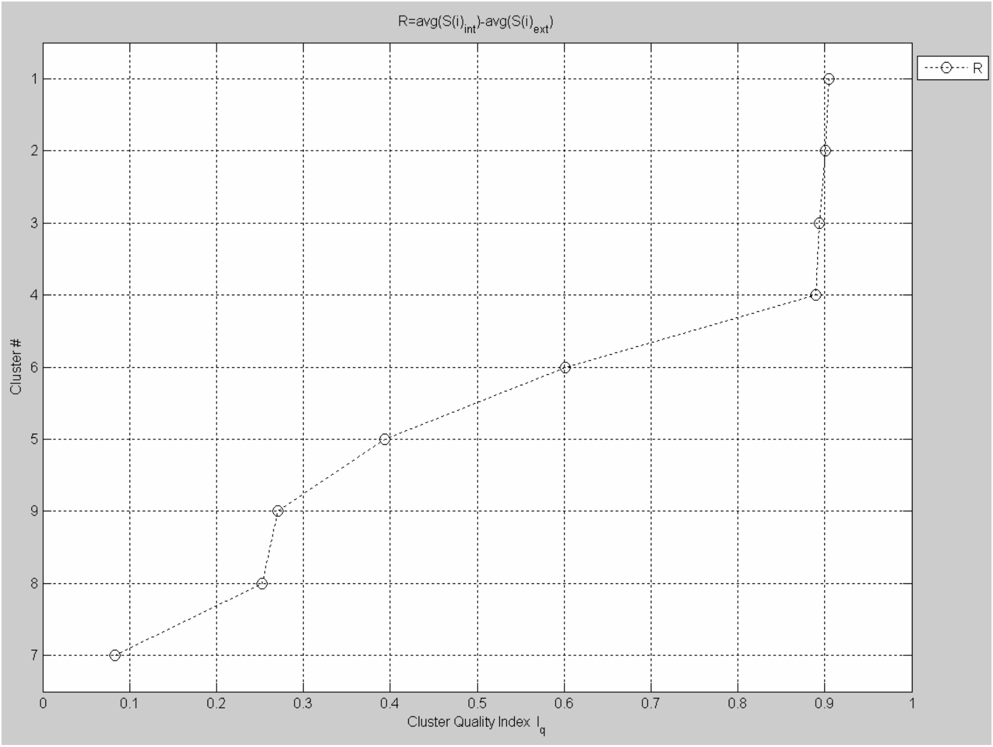

(a)

(b)

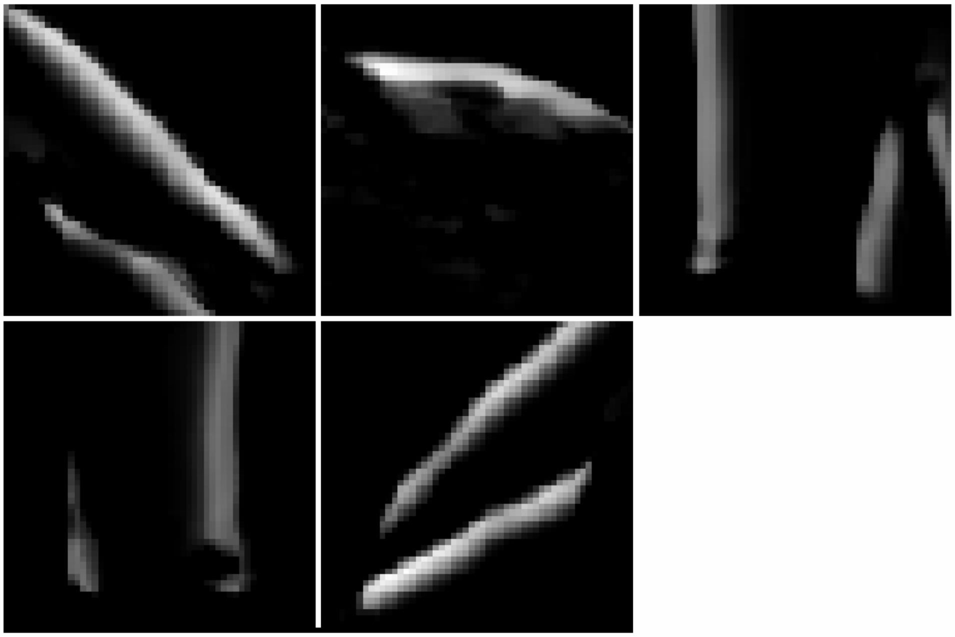
Figure 4.9 Taking the image sequence consisting of 120 rotated doll images (177×314 pixels) as input, the learned parts for each image at least include two arms and two legs.


0.23<s_ij≤ 0.61
0.61<s_ij≤ 0.82
0.82<s_ij≤ 1.00

(a)

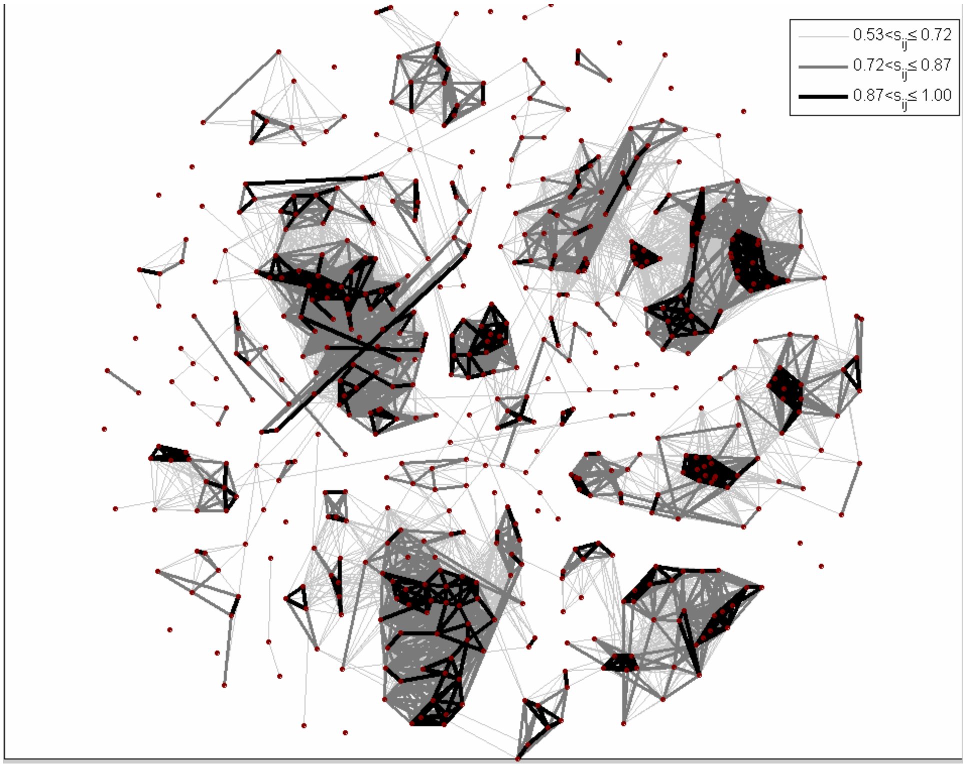

R=avg(S(i)_int)-avg(S(i)_ext)
R
Cluster #
1
2
3
4
6
5
9
8
7
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Cluster Quality Index I_q

(b)



(c)

Figure 4.10Over 30 frames, the experimental results are shown as: (a) the clusters grouped by the temporal SOM method; (b) the quality index I_q in rank order for the clusters; (c) the selected “centroparts” associated to the most robust clusters.



(a)

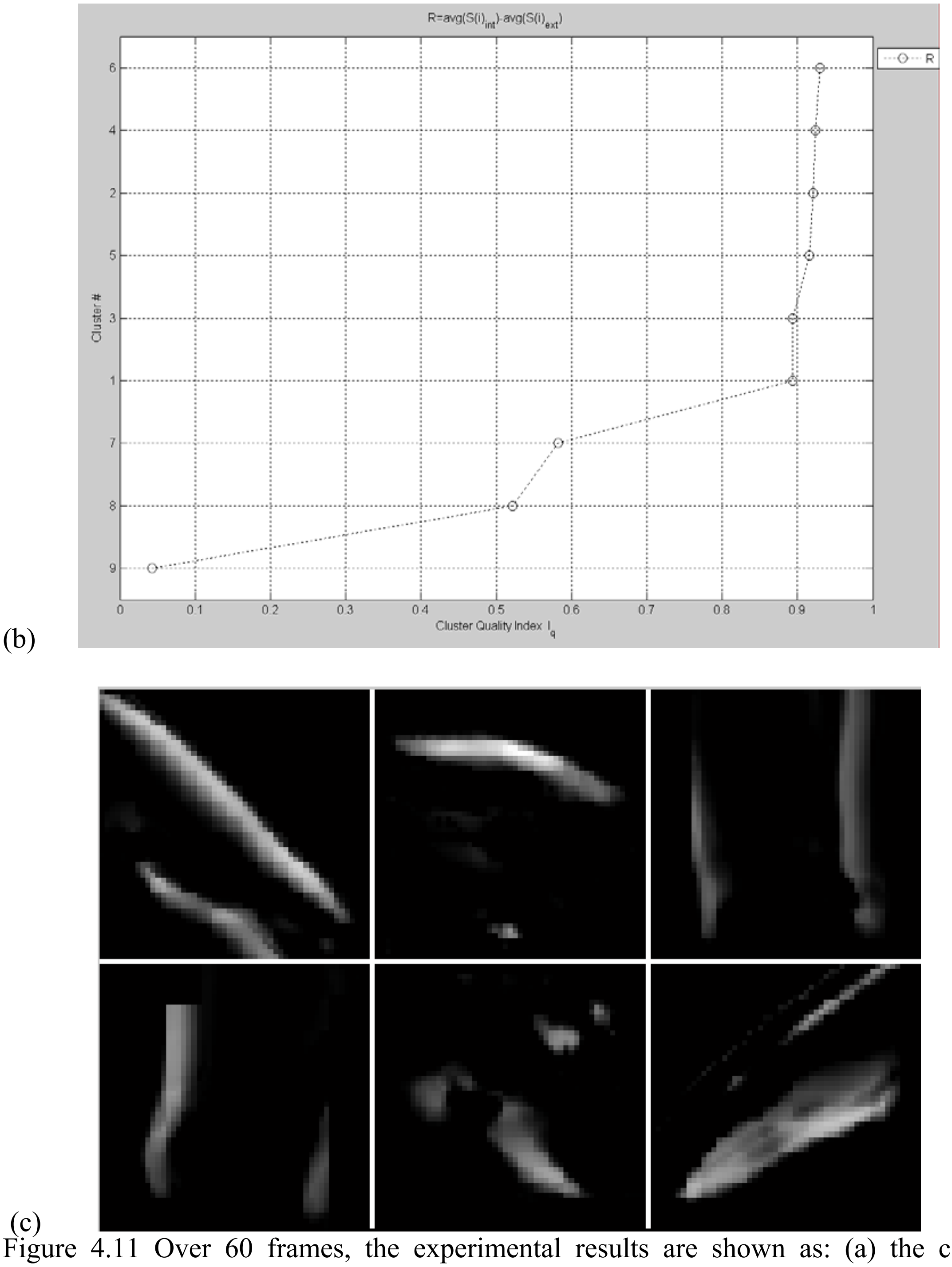


Figure 4.11 Over 60 frames, the experimental results are shown as: (a) the clusters grouped by the temporal SOM method; (b) the quality index I_q in rank order for the clusters; (c) the selected “centroparts” associated with the most robust clusters.

With the rotations of an object exceeding 90 degrees, the transformation of object’s shape increases. The quality of clusters grouped by above TKM method decreased (see

Fig. 4.12). In order to establish temporal associations between consecutive patterns in a temporal sequence, the network must be able to retain information about precious patterns in the sequence. This retention mechanism, usually called short-term short-term memory (STM), is then used to consider the temporal order and /or temporal dependencies between successive input samples. Here, we use the STM to store the "centroparts" selected over 30 continuous frames. After storing the intermediate results, the weights of the SOM layer are reset except the learned parts are continuously projected in the signal space. The procedure of evaluating the quality of cluster and selecting the "centroparts" are repeated over 120 frames. Finally, the intermediate results are merged based on similarity to minimize the "best" set of parts (see Fig. 4.13). This modification reduced the representative parts over 180º rotation angle from 1080 to 14.

Our experimental results show that the model learned and tested is not rotation invariant, because rotation causes significant differences between the object's actual shape and its projected images. However, we typically see objects from many different perspectives and manage to recognize them well despite the variations in appearance. Like size, shape is a property of objects that people usually perceive as constant despite

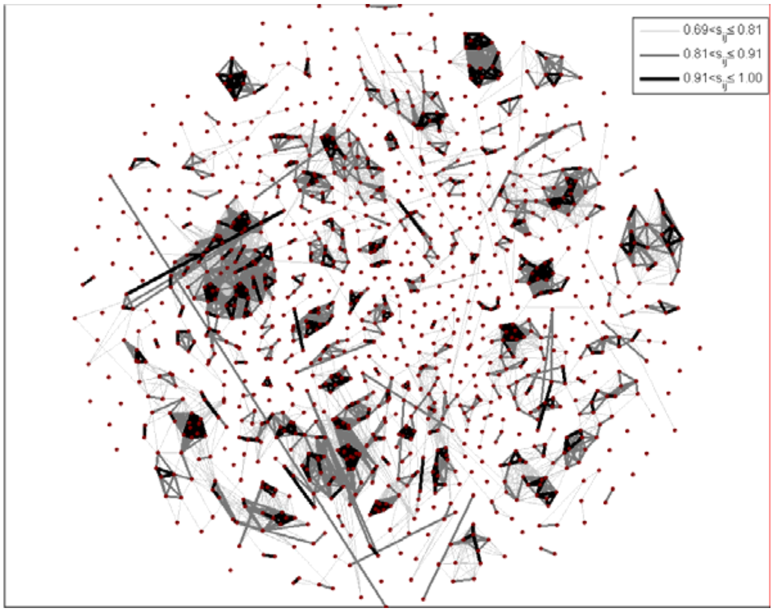

$0.69 < s_{ij} \leq 0.81$
$0.81 < s_{ij} \leq 0.91$
$0.91 < s_{ij} \leq 1.00$

Figure 4.12 Over 120 frames (0° ~ 180°), the quality of the clusters decreased.

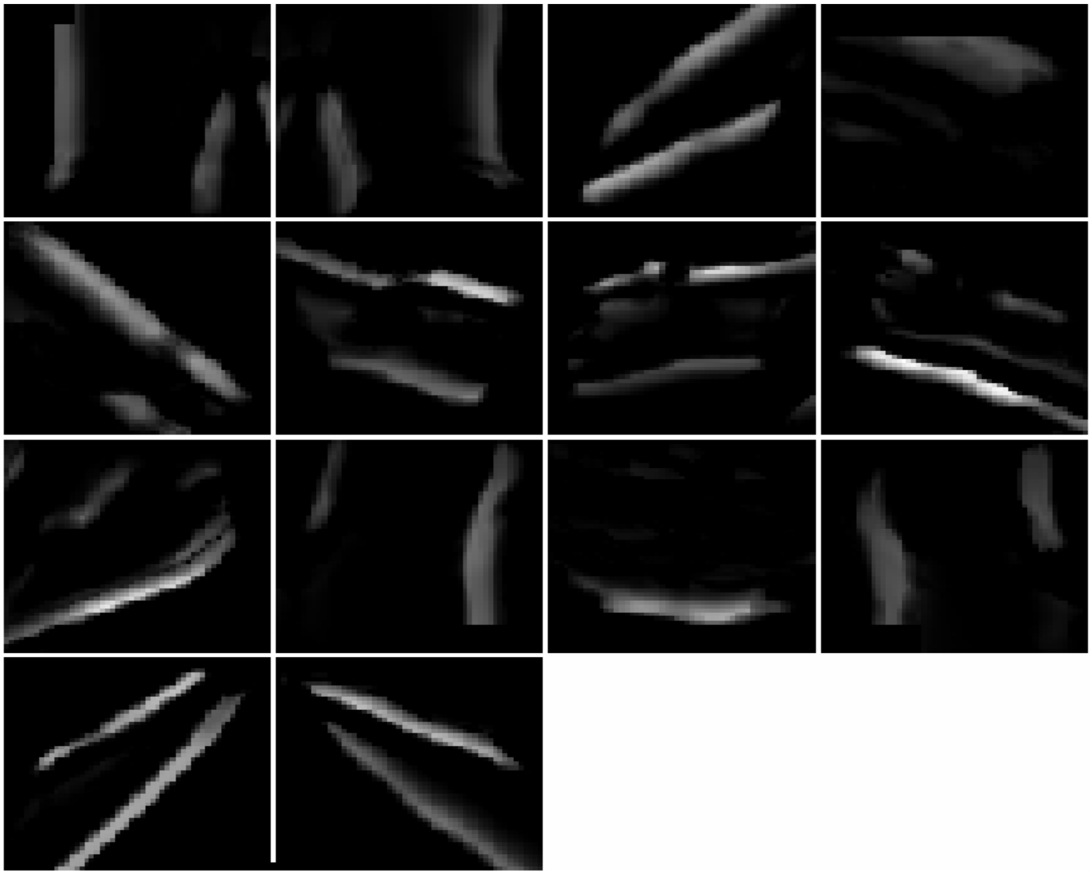


Figure 4.13 The representative parts of a doll object over 180º rotation angles.

changes in viewing perspective. Our intuitions about such shape constancy are based primarily on situations in which we perceive the same object from different view points by continuously moving from one view to another. Under these conditions, the perspective changes from moment to moment are gradual and easily perceived as shape preserving, especially when the object is being viewed continuously.

In a normal visual experience, different views of an object tend to appear in close temporal sequence as an animal manipulates the object or navigates around it. Similarly, the image sequence caught by a camera more or less varies smoothly over time. To take advantage of these general properties of the physical world, our challenge is in developing viewpoint invariant representations of object shapes by capturing the temporal relationships among the input image sequences.

As the spin-off of the reliability analysis of learned parts, we have demonstrated that our methods can successfully compressed the representative parts of the same object over

180º of rotation. It is easy to store the representative parts of rotation objects in long-term memory and retrieve them as templates for matching to new learned parts, which can achieve approximately invariant object recognition in certain simple cases. Moreover, we need quantify the utility of the learned RFs/parts for supporting visual tasks, such as object recognition in IT area.

Chapter 5

Unsupervised Object Recognition

5.1 Object Recognition

Visual object recognition, a sub-discipline of vision, addresses the problems of classifying and identifying objects in images. Human visual recognition is characterized by an ability to recognize novel objects for which the observer has no specific prior model. A number of factors contribute to the difficulty of object recognition tasks (Buxton and Neumann 1996). One is the complexity of the scene, which includes the number of objects in the image, the presence of objects that are touching and partly occluding one another, backgrounds that are highly textured or cluttered, and poor lighting conditions. Another is the generality of the object models. The third is the number of object models that a recognition system must consider. And the fourth is the complexity of the viewing transformation that maps the model coordinate frame to the image coordinate frame. Therefore, an adequate approach of object recognition must account for: (a) the accuracy of object recognition over changes in object size, location, and orientation; (b) the means by which the spatial relations between the parts or features of an object are represented; and (c) the attributes of both basic-level and subordinate-level recognition (Edelman 1997).

Feature-based approaches to object recognition generally operate by recovering a correspondence between local attributes or features of an image and an object model. The features are usually geometrical, and are often based on detecting intensity edges (places where there is a large change in image brightness) in the image. Brightness changes often correspond to the boundaries of objects or to surface markings on the objects. Local geometrical features can be simple, like corners, or involve a more complex fitting of geometrical primitives, such as quadratic curves. View-based approaches learn a model of the object's appearance in a two-dimensional image under different poses and illumination conditions. Often, each image is represented by a raster-scan of pixels, i.e., a vector of intensity values. At evaluation time, given a two-dimensional image the learned model is used to determine if the target object is present in the image or not. The most successful view-based approaches to object recognition are based on subspace techniques, which use principal components (or eigenvector) analysis to produce keys and form a concise description of a given set of images. The main advantage of such methods is that they are useful for tasks in which there is a large database of objects to be searched. The main disadvantage is that in general they do not work well with occlusion or with complex scenes and cluttered backgrounds, because the measure of similarity is sensitive to such variation.

The appearance of objects is diverse and complex. Models that are able to represent categories as diverse as dolls, fighters, snakes, clocks need to incorporate hundreds, if not thousands of parameters. A well-known rule-of-thumb says that the number of training examples has to be 5 to 10 times the number of object parameters, requiring a large training set. The penalty for using small training set is over fitting: while in-sample performance may be excellent, generalization to new examples is poor. Consequently, current systems are impractical where real-time user interaction is required. By contrast, humans can learn a new category quickly and easily, generally only requiring a few training examples. For instance, given 2 or 3 images of an animal you have never seen before, you can usually recognize it reliably later on. Moreover, we typically see objects from many different perspectives and manage to recognize them reasonably well despite variations in appearance. For example, we can recognize a specific face among many, despite change in viewpoint, scale, illumination or expression (though not rotation). This is to be contrasted with the state of the art in computer vision, where object recognition is typically not invariant in normal conditions.

In the precious chapters, we explored this human efficiency in a “ventral pathway”-like hierarchical network with sparse coding constraints. Here, we are interested in the problem of learning object classes from unsupervised visual experience. The main challenge is classifying the target objects independently of their rotation around the horizontal axis. In this case, the object shapes are partially occluded.

5.2 Object Recognition in Inferior Temporal (IT) Cortex

Visual object recognition is fundamental to the behavior of higher primates. It is also the most remarkable achievement of the visual cortex and one that probably influences its functional architecture most. The visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes – an extraordinarily difficult computational task.

During the past decade, there have been major advances in our understanding of how object recognition is performed in the primate visual system. There is now a broad body of evidence (Saleem, Cheng et al. 1995; Tanaka 1996; Booth and Rolls 1998; Kobatake, Wang et al. 1998; Haxby, Gobbini et al. 2001; Rolls, Aggelopoulos et al. 2003) showing that object recognition makes use of cells in inferior temporal cortex (IT) that respond to features of intermediate complexity. These features are typically invariant to a wide range of changes in location, scale, and illumination, while being sensitive to particular combinations of local shape, color, and texture properties.

An understanding of the intermediate-level features in visual cortex was first obtained in a novel approach developed by Tanaka and his associates (Tanaka 1993; Kobatake and Tanaka 1994). They recorded from individual cells in anesthetized monkeys while using a library of objects and a computer graphics editing system to characterize the cell responses of monkeys. First, a collection of complex real-world objects were presented to the monkeys to obtain an initial response. An image of the best initial object was then subject to numerous attempts at simplification and modification to obtain an optimal response. Although some cells in anterior IT cortex responded to simple line or bar features, in most cases the optimal response was obtained by features of intermediate complexity, such as a dark five-sided star shape, a circle with a thin protruding element at a particular orientation, or a green horizontal textured region within a triangular boundary. Some cells responded only to more complex shapes, such as moderately detailed face or hand images (see Fig. 5.1). Some cells respond similarly to different views of the same object. These cells could respond to differently shaped images, with different features, when they were part of one object, and might not respond to superficially similar images and features when they were part of another object. This evidence suggests that these cells are not simply feature detectors, responding to a particular feature found in some images, but rather are responding to the object(s) per se.

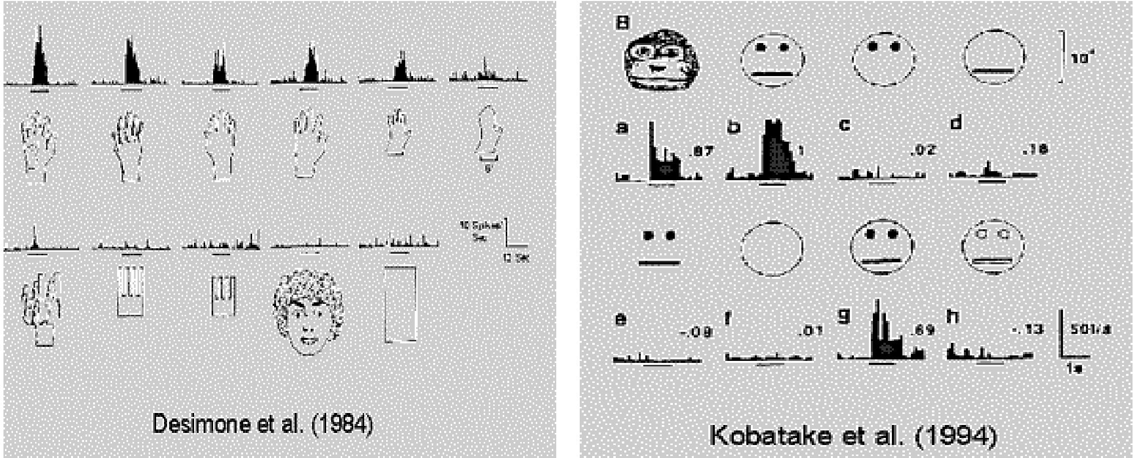


Figure 5.1 Some cells in IT responded only to more complex shapes, such as moderately detailed face or hand images.

These intermediate-complexity cells were often highly sensitive to small variations in shape, such as the degree of rounding of corners, or relative lengths of elements. On the other hand, the cells exhibited a wide range of invariance to other parameters, such as retinal location, size, and contrast. Detailed studies of size and position invariance (Ito, Tamura et al. 1994; Booth and Rolls 1998) have shown that about 55% of the cells have a size invariance range of greater than 2 octaves and 20% have a range of more than 4 octaves. Most cells have a receptive field covering a large portion of the image (an average 25 degrees of visual angle, usually including all of the fovea). These properties would seem ideally suited to determining object identity without needing to replicate the cell for each combination of values of the image parameters.

Some IT cells of macaques have responses that provide information about objects or faces. The responses of these cells are often relatively invariant with respect to the position in the visual field, size, and even view of the object. The firing rate differences between different objects and the firing rate similarities within objects (in different views) lead us to believe that visual features are not responsible for these invariant responses, and that the cells are responding when a certain object is being seen. The anatomical location of the view-invariant cells suggests that these cells do not form a separate anatomical population of cells in IT, but rather are intermingled with cells that are view-dependent and require a certain feature or combination of features for activation. This supports the notion that these view-invariant responses are being formed by associating together the responses of view-dependent visual cells. We note that view-invariant representations of objects by cells in brain areas such as the IT would be computationally advantageous, in that subsequent processing structures (such as the hippocampus and the amygdala) that receive inputs from IT cortex would not need to learn about all possible sizes, positions and views of each object, and there would be easy generalization between one size or position of an object and other sizes, positions or views of the same object.

Cells that were close together in cortex often responded to small variations of the same feature. Based on the average size of these related feature columns as a proportion of the total size of the brain region, Tanaka (Tanaka 1996) estimated that there was room for about 1300 such feature columns (see Fig. 5.2). However, if the column sizes vary, then the large ones would be preferentially sampled and the total number of feature types could be far greater.

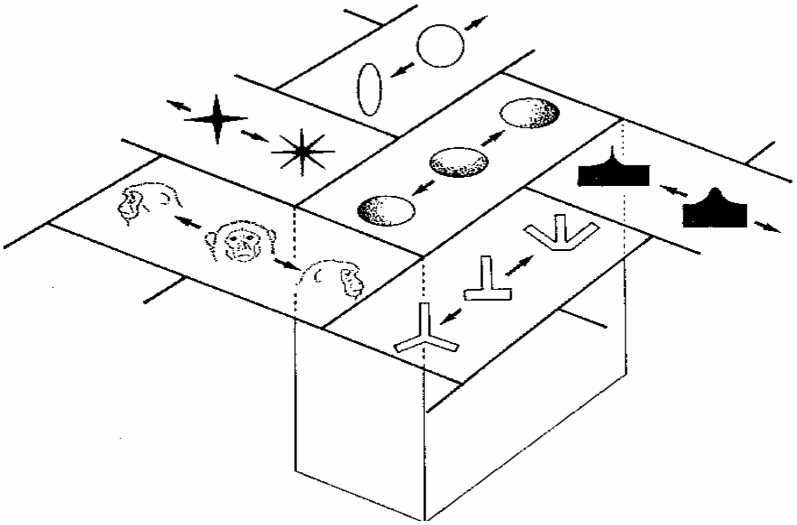


Figure 5.2 Neighboring columns in IT have cells that respond to stimuli that share similar feature (Tanaka 1996).

The feature responses have been shown to depend on previous visual learning during exposure to specific objects containing the features. (Logothetis, Pauls et al. 1995) examined the responses of cells in monkeys that had been trained to classify views of wire-frame and spheroidal shapes. They discovered many cells that responded only to particular views of these shapes, while exhibiting the usual invariance to large ranges of scale and location. (Booth and Rolls 1998) found similar results for 10 plastic objects that had been placed in the monkey's cage for a period of weeks or months without any training. In addition to the usual view-dependent cells, they found a small population of cells that responded to any view of a particular object (these responded to a conjunction of shape views, rather than to a simple feature such as color that was shared between views). In a dramatic illustration of learning, (Tovee, Rolls et al. 1996) showed that a face sensitive cell could learn to recognize degraded face images (that were previously unrecognizable) by exposure to 5 seconds of training images that showed the transition between normal and degraded images.

It is also known that object recognition in human vision uses a process of serial attention to bind features to object interpretations, determine pose, and segment an object from a cluttered background (Wolfe and Bennett 1997; Reynolds and Desimone 1999). A wide range of psychophysical experiments have shown that pre-attentive object descriptions consist of a collection of isolated features, and that serial attention is necessary to represent shape relationships and integrate features into a common object description.

There has been some previous research (Mel 1997; Lowe 2000) on building computer systems for object recognition that use intermediate features similar to those in IT cortex. One problem has been that these earlier systems use correlation to estimate the presence of intermediate features in an image, which has a prohibitive computational cost due to the need to compare each feature at every location, scale, and orientation to the image. In this chapter, we present a method to learn object class models for object recogntion from different viewing angles. We focus on a model where objects are represented as flexible constellations of rigid parts. The variability within a class is represented by a joint probability density function (pdf) on the shape of the constellation. In the first stage, the

V4 parts decomposition model (see Chapter 4 for more detail) automatically identifies distinctive parts of object in the training set. The set of model parameters, including the shape pdf, is then learned using the expectation maximization (EM) algorithm. The method achieves very good classification results with approximate viewpoint invariance.

5.3 Baye Theorem

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

$P(h)$ is often called the prior probability of h and may reflect any background knowledge we have about the chance that h is a correct hypothesis. If we have no such prior knowledge, then we might simply assign the same prior probability to each candidate hypothesis. Similarly, we will write $P(D)$ to denote the prior probability that training data D will be observed (i.e. the probability of D given no knowledge about which hypothesis holds). Next, we will write $P(D/h)$ to denote the probability of observing data D given some world in which hypothesis h holds. $P(h/D)$ is called the posterior probability of h, because it reflects our confidence that h holds after we have seen the training data D. Notice the posterior probability $P(h/D)$ reflects the influence of the training data D, in contrast to the prior probability $P(h)$, which is independent of D. Bayes theorem:

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)} \tag{5.1}$$

is cornerstone of Bayesian learning methods because it provides a way to calculate the posterior probability $P(h/D)$, from the prior probability $P(h)$, together with $P(D)$ and $P(D/h)$.

As one might intuitively expect, $P(h/D)$ increases with $P(h)$ and with $P(D/h)$ according to Bayes theorem. It is also reasonable to see that $P(h/D)$ decreases as $P(D)$ increases, because the more probable it is that D provides in support of h.

In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis. We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis. More precisely, we will say that h_{MAP} is a MAP hypothesis provided

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P\left(h/D\right) \\ &= \arg\max_{h \in H} \frac{P\left(D/h\right) p(h)}{P(D)} \\ &= \arg\max_{h \in H} P\left(D/h\right) P(h) \qquad (5.2) \end{aligned}$$

In the final step above we dropped the term $P(D)$ because it is a constant independent of h.

In some cases, we will assume that every hypothesis in H is equally probable a priori (($P(h_i) = P(h_j)$ for all h_i and h_j in H). In this case w can further simplify Eq. (5.2) and need only consider the term $P\left(D/h\right)$ to find the most probable hypothesis. $P\left(D/h\right)$ is oftern called the likelihood of the data D given h, and any hypothesis that maximizes $P\left(D/h\right)$ is called a maximum likelihood (ML) hypothesis, h_{ML}

$$h_{ML} = \arg\max_{h \in H} P\left(D/h\right) \qquad (5.3)$$

5.3.1 The EM Algorithm

In many practical settings, some of the relevant instance variables might be unobservable. The EM algorithm (Dempster, Laird et al. 1977) provides a general approach to learning in the presence of unobservable variables. This algorithm begins with an arbitrary initial hypothesis. It then repeatedly calculates the expected values of the hidden variables (assuming the current hypothesis is correct), and then recalculates the maximum likelihood hypothesis (assuming the hidden variables have the expected values calculated by the first step). This procedure converges to a local maximum likelihood hypothesis, along with estimated values of the hidden variables.

The EM algorithm can be applied in many settings where we wish to estimate some set of parameters θ that describe an underlying probability distribution. In general let $X = \{x_1, \cdots x_m\}$ denote the observed data in a set of m independently drawn instances, let $Z = \{z_1, \cdots, z_m\}$ denote the unobserved data in these same instances, and let $Y = X \cup Z$ denote the full data. Note the unobserved Z can be treated as a random variable whose probability distribution depends on the unknown parameters θ and on the observed data X. Similarly, Y is a random variable because it is defined in terms of the random variable Z. We use h to denote the current hypothesized values of the parameters θ, and h' to denote the revised hypothesis that is estimated on each iteration of the EM algorithm.

The EM algorithm searches for the maximum likelihood hypothesis h' by seeking h' that maximizes $E\left[\ln P(Y/h')\right]$. This expected value is taken over the probability distribution governing Y, which is determined by the unknown parameters θ. Let us consider exactly what this expression signifies. First, $P(Y/h')$ is the likelihood of the full data Y given hypothesis h'. It is reasonable that we want to find a h' that maximizes some function of this quantity. Second, maximizing the logarithm of this quantity, $\ln P(Y/h')$, also maximizes $P(Y/h')$. Third, we introduce the expected value $E\left[\ln P(Y/h')\right]$ because the full data Y is a combination of the observed data X and unobserved data Z, we must average over the possible values of the unobserved Z, weighting each according to its probability. In other words we take the expected value $E\left[\ln P(Y/h')\right]$ over the probability distribution governing the random variable Y. The distribution governing Y is determined by the completely known values for X, plus the distribution governing Z.

What is the probability distribution governing Y? In general we will not know this distribution because it is determined by the parameters θ that we are trying to estimate. Therefore, the EM algorithm uses its current hypothesis h in place of the actual parameters θ to estimate the distribution governing Y. Let us define a function $Q\left(h'/h\right)$ that gives $E\left[\ln P(Y/h')\right]$ as a function of h', under the assumption that $\theta = h$ and given the observed portion X of the full data Y.

$$Q\left(h'/h\right) = E\left[\ln P(Y/h')/h, X\right] \tag{5.4}$$

We write this function Q in the form $Q\left(h'/h\right)$ to indicate that it is defined in part by the assumption that the current hypothesis h is equal to θ. In its general form, the EM algorithm repeats the following two steps until convergence:

Step 1: *Estimation (E) step*: Calculate $Q\left(h'/h\right)$ using the current hypothesis h and the observed data X to estimate the probability distribution over Y.

$$Q\left(h'/h\right) \leftarrow E\left[\ln P(Y/h')/h, X\right] \tag{5.5}$$

Step 2: *Maximization (M) step:* Replace hypothesis h by the hypothesis h' that maximizes this Q function.

$$h \leftarrow \underset{h'}{\arg\max}\, Q\left(h'/h\right) \tag{5.6}$$

When the function Q is continuous, the EM algorithm converges to a stationary point of the likelihood function $P(Y/h')$. When this likelihood function has a single maximum, EM will converge to this global maximum likelihood estimate for h'. Otherwise, it is guaranteed only to converge to a local maximum. In this respect, EM shares some of the same limitations as other optimization methods such as gradient descent, line search, and conjugate gradient.

5.4 Model

We are interested in the problem of recognizing members of object classes, where we define an object class as a collection of objects which share characteristic features or parts that are visually similar and occur in similar spatial configurations. We model object classes following the work of (Burl, Leung et al. 1996). An object is composed of parts, where ‘parts’ are image patches that may be detected and characterized by appropriate detectors. Shape describes the geometry of the mutual position of the parts in a way that is invariant with respect to rigid and, possibly, affine transformations (Leung, Burl et al. 1998). When building models for object classes of this type, one is faced with three problems. 1) *Segmentation or registration of training images*: Which objects are to be recognized and where do they appear in the training images? 2) *Part selection*: Which object parts are distinctive and stable? 3) *Estimation of model parameters*: What are the parameters of the global geometry or shape and of the appearance of the individual parts that best describe the training data?

Although solutions to the model learning problem have been proposed (Amit and Geman 1999; Schneiderman and Kanade 2000; Viola and Jones 2001), they typically require that one of the first two problems, if not both, be answered by a human supervisor. For example, features in training images might need to be hand-labeled. Training images showing object in front of a uniform background may be required. Or objects might need to be positioned in the same way throughout the training images so that a common reference frame can be established.

The problem of automatic *part selection* is important, since it is generally not established that parts that appear distinct to the human observer will also lend themselves to successful detection by a machine. We introduced a completely unsupervised solution of decomposing and selecting distinctive and stable parts in chapter 4. In this chapter, a solution to the estimation of model parameters is provided by the EM framework, which allows simultaneous estimation of unknown data and their probabilities.

We use a joint probability density on part appearance and shape to model the object class. Object classification is performed by first running part decomposition (Yang and Jabri 2006) on the image, thus obtaining a set of candidate parts , which include part shape and part position information. The second stage consists of forming likely object hypotheses, i.e. constellations of appropriate parts (e.g. arms, legs and head); both complete and partial constellations are considered, in order to allow for partial occlusion or deformation caused by rotation. The third stage consists of using the object's joint probability density for either calculating the likelihood that any hypothesis arises from an object (object detection), or the likelihood that one specific hypothesis arises from an object (object localization). In order to train a model we need to decide on the key parts of the object, select corresponding parts (e.g. arms, legs etc) on multiple-view images of an object, and lastly we need to estimate the joint probability density function on part appearance and shape. The method we present here performs all three stages automatically. A block diagram is shown in Fig 5.3.

We first use the V4 part decomposition model to detect potentially informative parts by sparse and unsupervised learning algorithm (see Chapter 4). The second step of our proposed model learning algorithm simultaneously estimates which features actually are the most informative, and probability of the constellations that they tend to form when they are associated to an object of interest. This process requires iterating four operations: (a) choosing a tentative model structure, i.e. the collection of features (or parts) that are associated to the object, (b) establishing a correspondence between homologous parts across the training set, and simultaneously labeling as 'background' or 'noise' all parts that are not put in such correspondence, (c) estimating the joint model probability density from such a labeled training set, (d) assessing the performance of such a model. In our model operations (b) and (c) are performed only implicitly in a "soft" way, using expectation maximization (EM). At each iteration, the parameters of the underlying probability model are estimated. Depending on the performance of the model on a validation data set, the choice of parts is modified. This process is iterated until the final model is obtained where no further improvements are possible.

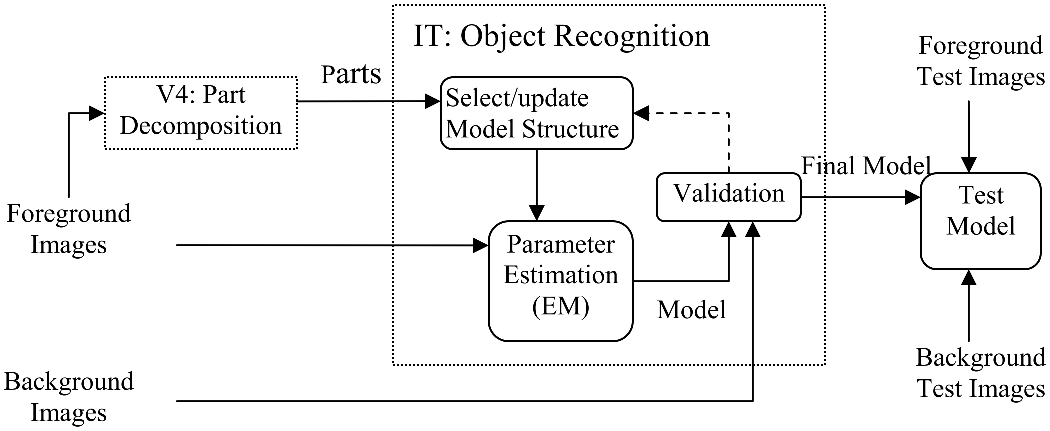


Figure 5.3 Block diagram of the method for unsupervised object recognition. "Foreground images" are images containing the target objects. "Background images" do not contain the target objects. (Inspired by Webber and Welling 2000)

5.4.1 Modeling Objects in Images

We model objects as collections of rigid parts. The part detection stage (see chapter 4) therefore transforms an entire image into a collection of parts. Some of those parts might correspond to an instance of the target object class (the *foreground*), while others stem from background clutter or are simply false detections (the *background*). Throughout this chapter, the only information associated with an object part is its position in the image and its identity or part type. We assume that there are T different types of parts. The positions of all parts extracted from one image can be summarized in a matrix-like form,

$$X^o = \begin{bmatrix} x_{11}x_{12} & \cdots & x_1N_1 \\ x_{21}x_{22} & \cdots & x_2N_2 \\ & \vdots & \\ x_{T1}x_{T2} & \cdots & x_TN_T \end{bmatrix} \tag{5.7}$$

where the superscript 'o' indicates that these positions are observable in an image, as opposed to being unobservable or missing, which will be denoted by 'm'. Thus, the t^{th} row contains the locations of detections of part type t, where every entry, x_{ij}, is a two dimensional vector. If we now assume that an object is composed of F different parts, we need to be able to indicate which parts in X^o correspond to the foreground (the object of interest). For this we use the vector **h**, a set of indices, with $h_i = j, j > 0,$ indicating that point x_{ij} is a foreground point. If an object part is not contained in X^o, because it is occluded or otherwise undetected, the corresponding entry in **h** will be zero. Often, we do not know which parts correspond to the foreground and which parts correspond to the background (noise). Therefore, **h** is not observable and we will treat it as hidden or missing data. We call **h** a *hypothesis*, since we will use it to hypothesize that certain parts of X^o belong to the foreground object. It is also convenient to represent the positions of any unobserved object parts in a separate vector X^m which is, of course, hidden as well. The dimension of X^m will vary, depending on the number of missed parts.

We can now define a generative probabilistic model through the joint probability density

$$p(X^o, X^m, h) \tag{5.8}$$

Note that not only the entries of X^o and X^m are random variables, but also their dimensions.

In order to provide a detailed parameterization of Eq. (5.8), we introduce two auxiliary variables, **b** and **n**. The binary vector **b** encodes information about which parts have been detected and which have been missed or occluded/deformed. Hence, $b_f = 1$ if $h_f > 0$ and $b_f = 0$ otherwise. The variable **n** is also a vector, where n_t shall denote the number of *background (or noise)* candidates included in the t^{th} row of X^o . Since both variables are completely determined by **h** and the size of X^o, we have $p(X^o, X^m, h) = p(X^o, X^m, h, n, b)$. Since we assume independence between foreground and background, and, thus, between $p(n)$ and $p(b)$, we decompose in the following way

$$p(X^o, X^m, h, n, b) = p(X^o, X^m \mid h, n) p(h \mid n, b) p(n) p(b) \tag{5.9}$$

The probability density over the number of background detections can be modeled by a Poisson distribution,

$$p(n) = \prod_{t=1}^{T} \frac{1}{n_t!} (M_t)^{n_t} e^{-M_t} \tag{5.10}$$

where M_t is the average number of background detections of type t per image. This conveys the assumption of independence between part types in the background and the idea that background detections can arise at any location in the image with equal probability, independently of other locations. For a discrete grid of pixels, $p(n)$ should be modeled as a binomial distribution. However, since we will model the foreground detections over a continuous range of positions, we chose the Poisson distribution, which can be derived as a limiting case of the binomial distribution. Admitting a different M_f for every part type allows us to model the different detector statistics.

The probability $p(b)$ is modeled explicitly by a table of size 2^F which equals the number of possible binary vectors of length F . If F is large, the explicit probability table might become unreasonably large. In this case we can assume independence between features/parts and model $p(b)$ by a product of independent densities, $p(b)=\prod_{f=1}^{F} p(b_f)$. The number of parameters is reduced in that case from 2^F to F.

The density $p(h \mid n, b)$ is modeled by,

$$p(h \mid n,b)=\begin{cases} \dfrac{1}{\prod_{f=1}^{F} N_f^{\;b_f}} & h \in H(b,n) \\ 0 & other \quad h \end{cases} \tag{5.11}$$

where $H(b,n)$ denotes the set of all hypotheses consistent with **b** and **n**, and N_f denotes the total number of detections of the type of part f . This expresses the fact that all consistent hypotheses, the number of which is $\prod_{f=1}^{F} N_f^{\;b_f}$, are equally likely in the absence of information on the part locations.

Finally, we use

$$p(X^o, X^m \mid h,n) = p_{fg}(X^o, X^m) p_{bg}(x_{bg}) \tag{5.12}$$

where we defined $z^T = (X^{oT} X^{mT})$ as the coordinates of all foreground detections (observed and missing) and x_{bg} as the coordinates of all background detections. Here we have made the important assumption that the foreground detections are independent of the background. In our experiments, $p_{fg}(z)$ is modeled as a joint Gaussian with mean μ and covariance Φ. So far, we have modeled only absolute part positions in the image. This is of little use, unless the foreground object is in the same positions in every image. We can, however, obtain a translation invariant formulation of our algorithm as discussed in chapter 4 by describing all part positions relative to the position of one reference part. Under this modification, p_{fg} will remain a Gaussian density, and therefore not introduce any fundamental difficulties. The positions of the background detections are modeled by a uniform density,

$$p_{bg}(X_{bg}) = \prod_{f=1}^{F} \frac{1}{A^{n_f}} \tag{5.13}$$

where A is the total image area.

5.4.2 Classification

Throughout the experiments presented here, our objective is to classify images into the classes "object present" (class C_1) and "object absent" (class C_2). Given the observed data, $p(X^o)$, the optimal decision—minimizing the expected total classification error--is made by choosing the class with maximum a posteriori probability (MAP approach). It is therefore convenient to consider the following ratio

$$\frac{p(C_1 \mid X^o)}{p(C_0 \mid X^o)} \infty \frac{\sum_h p(X^o, h \mid C_1)}{p(X^o, h_0 \mid C_0)} \tag{5.14}$$

where h_o denotes the *null hypothesis* which explains all parts as background noise. For convenience, we omitted the variables **b**=*sign*(**h**) (with *sign*(**0**)=0) and **n**=N-**b**. Notice that the ratio $\frac{p(C_1)}{p(C_0)}$ is omitted, since it can be absorbed into a decision threshold. The sum in the numerator includes all hypotheses, including the null hypothesis, since the object could be present but remain undetected by any part detector. In the denominator, the only consistent hypothesis to explain "object absent" is the null hypothesis. Although we are here concerned with classification only, our framework is by no means restricted to this problem. For instance, object localization is possible by identifying those foreground parts in an image, which have the highest probability of corresponding to an occurrence of the target object.

5.4.3 Model Learning

In order to train an object model on a set of images, we need to solve two problems. Firstly, we need to decide on the subset of the selected part candidates that will be used in the model, i.e. which define the model configuration. Secondly, we need to learn the parameters underlying the probability densities. We solve the first problem using an iterative, "greedy" strategy, under which we try different configurations. At each iteration, the pdfs are estimated using EM.

We now address the problem of estimating the model pdfs with a given set of model parts, from a set of N training images. Since our detection method relies on the MAP principle, it is our goal to model the class conditional densities as accurately as possible.

We therefore employ the EM algorithm to produce maximum likelihood estimates of the model parameters, $\theta = \{\mu, \Phi, p(b), M\}$. EM is well suited for our problem, since the variables **h** and X^m are missing and must be inferred from the observed data X^o. In the following we omit reference to the variables **n** and **b** because they are simple functions of **h**.

In standard EM fashion, we attempt to maximize the log-likelihood of the observed data, which is given as

$$L(X^o \mid \theta) = \sum_{i=1}^{N} \log \sum_{h_i} \int p(X_i^o, X_i^m, h_i \mid \theta) dX_i^m \tag{5.15}$$

Since maximizing sums and integrals of a logarithm is difficult in practice, we choose to iteratively optimize a sequence of cost functions—again in standard EM fashion:

$$Q(\theta_k \mid \theta_{k-1}) = \sum_{i=1}^{N} E_{k-1}\left[\log p(X_i^o, X_i^m, h_i \mid \theta_k)\right] \tag{5.16}$$

where k counts iterations and $E_{k-1}[\bullet]$ denotes expectation with respect to the posterior density $p(X_i^o, X_i^m, h_i \mid \theta_{k-1})$.

Formally, the E-step amounts to the calculation of this posterior density (or certain sufficient statistics thereof), while the M-step maximizes $Q(\theta_k \mid \theta_{k-1})$ over θ_k, given this posterior density with parameters from the previous iteration, θ_{k-1}. It can be shown that the EM algorithm converges to a local maximum of the log-likelihood.

We now derive update rules that are used in the M-step of the EM algorithm. The parameters we need to consider are those of the Gaussian governing the distribution of the foreground parts, i.e. μ and Φ, the table representing $p(b)$ and the parameter, M, governing the background densities. It is helpful to decompose Q into four parts, following the factorization in Eq. (5.9).

$$\begin{aligned}Q(\theta_k \mid \theta_{k-1}) &= Q_1(\theta_k \mid \theta_{k-1}) + Q_2(\theta_k \mid \theta_{k-1}) + Q_3(\theta_k \mid \theta_{k-1}) + Q_4 \\ &= \sum_{i=1}^{N} E_{k-1}[\log p(n_i \mid \theta_k)] + \sum_{i=1}^{N} E_{k-1}[\log p(b_i \mid \theta_k)] \\ &+ \sum_{i=1}^{N} E_{k-1}\left[\log p(X_i^o, X_i^m \mid h_i, n_i, \theta_k)\right] + \sum_{i=1}^{N} E_{k-1}[\log p(h_i \mid n_i, b_i)]\end{aligned} \tag{5.17}$$

Only the first three terms depend on parameters that will be updated during EM.

Taking the derivative of $Q(\theta_k \mid \theta_{k-1})$ with respect to $\mu, \Phi, p(b), M$ and equating this to zero produces the following update rules:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N} E_{k-1}\left[z_i\right] \tag{5.18}$$

$$\Phi_k = \frac{1}{N}\sum_{i=1}^{N} E_{k-1}\left[z_i z_i^T\right] - \mu_k \mu_k^T \tag{5.19}$$

$$p_k(\bar{b}) = \frac{1}{N}\sum_{i=1}^{N} E_{k-1}\left[\delta_b, \bar{b}\right] \tag{5.20}$$

$$M_k = \frac{1}{N}\sum_{i=1}^{N} E_{k-1}\left[n_i\right] \tag{5.21}$$

These update rules constitute the M-step. The E-step amounts to the calculation of$E\left[z_i\right]$, $E\left[z_i z_i\right]$, $E\left[\delta_b, \bar{b}\right]$ and $E\left[n_i\right]$. The posterior density is given by

$$p(h_i, X_i^m \mid X_i^o, \theta) = \frac{p(h_i, X_i^m, X_i^o \mid \theta)}{\sum_h \int p(h_i, X_i^m, X_i^o \mid \theta) dX_i^m} \tag{5.22}$$

where we omitted again the dependence on **b(h)** and **n(h)**. The denominator in the above expression, $p(X_i^o)$, is calculated as follows. Choose a hypothesis consistent with the observed data. The missing data in that hypothesis is eliminated by deleting the means and covariances of those dimensions of a Gaussian. Now repeat this operation, summing over all possible hypothesis. The expectations of the statistics are calculated in a similar fashion. $E\left[\delta_b, \bar{b}\right]$ is calculated by summing only over those hypotheses consistent with $\bar{b}$ in the numerator and dividing by $p(X_i^o)$. Similarly, $E\left[n_i\right]$ is calculated by averaging **n(h)** over all hypotheses. For $E\left[z\right] = \left(X^o \quad E\left[X^m\right]\right)$, we need

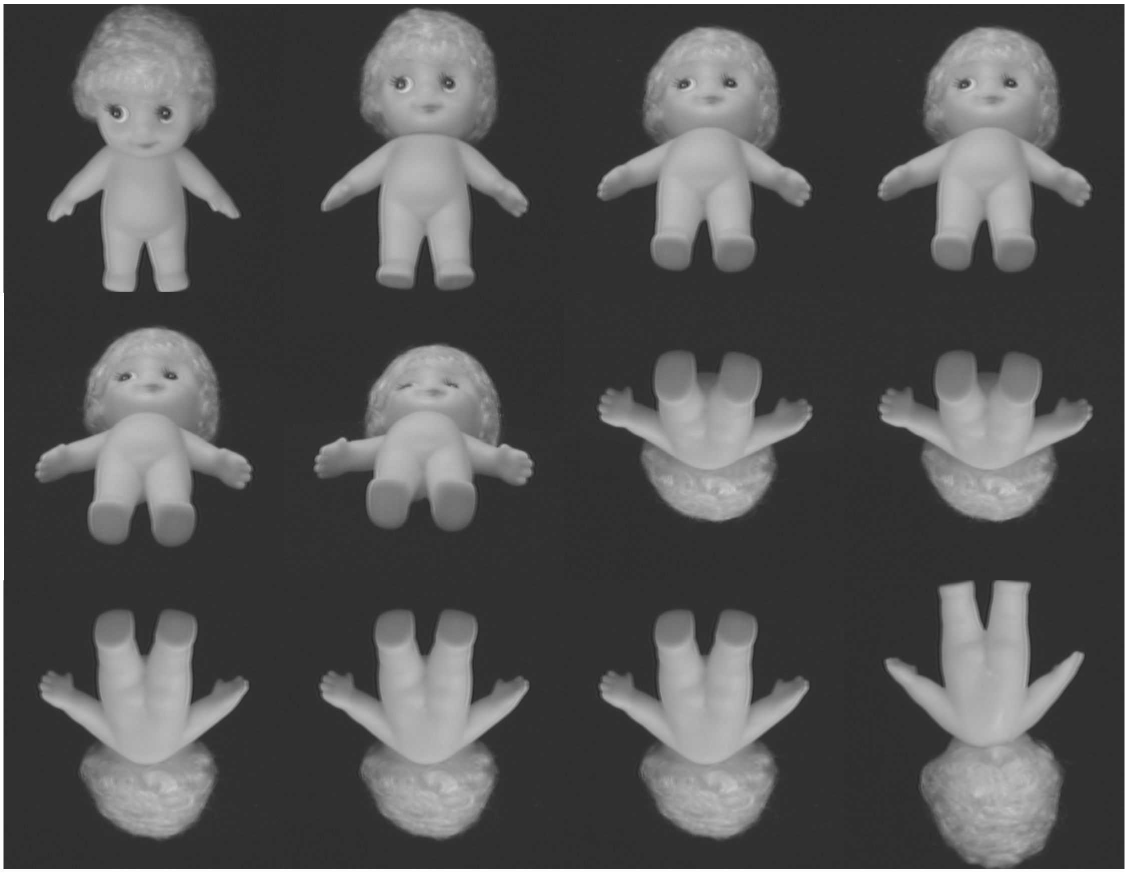
$$\int X^m G(z \mid \mu, \sigma) dX^m = \mu^m + \Phi^{mo} \Phi^{oo-1}(X^o - \mu^o) \tag{5.23}$$

where we defined $\mu = \left(\mu^o \quad \mu^m\right)$ and a similar decomposition for Φ. For the calculation of $E\left[z_i z_i\right]$, we need in addition to the above equation the following result

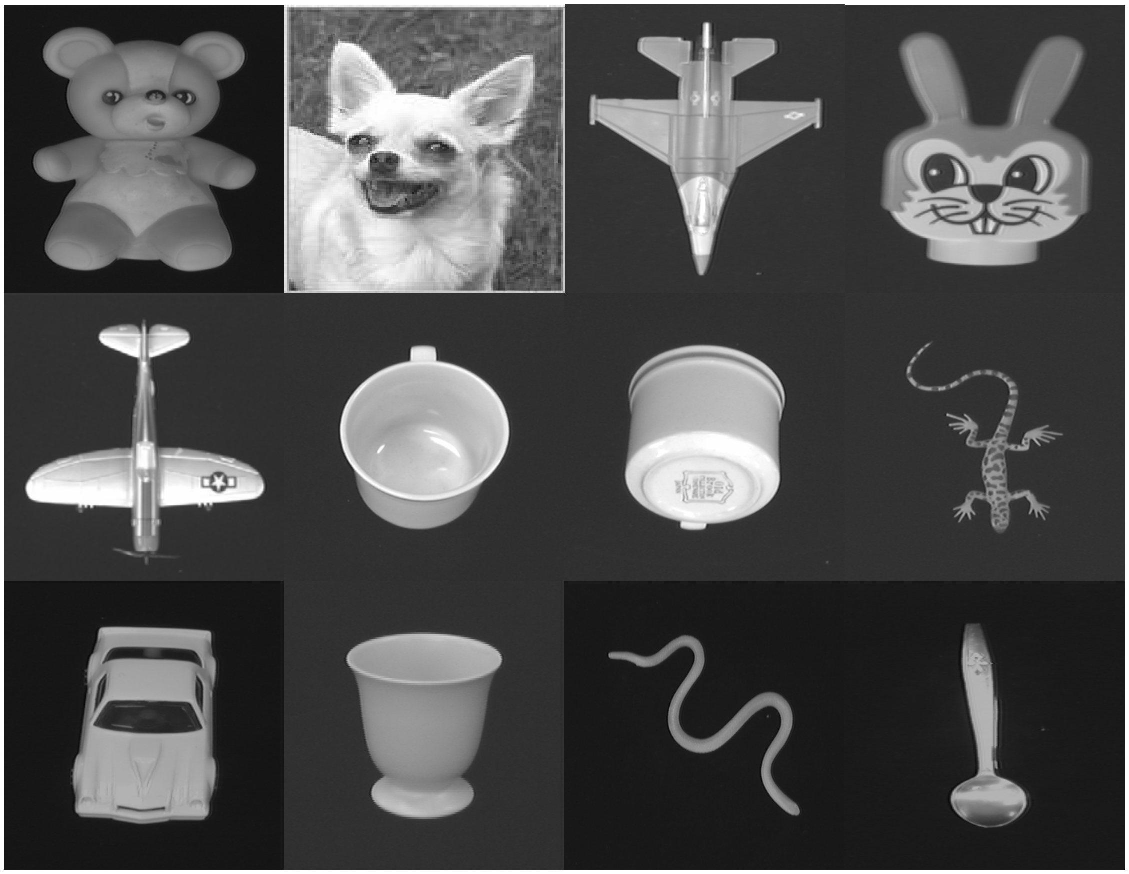
$$E\left[X^m X^{mT}\right] = \Phi^{mm} - \Phi^{mo}\Phi^{oo-1}\Phi^{moT} + E\left[X^m\right] E\left[X^m\right]^T \tag{5.24}$$

5.5 Results

The foreground images consisted of 120 doll images that rotated over up to 180º angle of rotation. This set was divided into two non-overlapping, equally large sets for training and testing. The set of background images contained different objects, such as fighters, cars, bears, and cups etc (see Fig 5.4). We initialized the model configuration with a small randomly sampled set and estimated the model parameters by running EM on the data. The number of features in all models was limited to four since we found that, due the limited number of training images, the learning algorithm was overfitting the training data when five or more features were used. Even with models with four features, the training error was often close to zero, while the test error was significantly larger.



(a) Foreground images



(b) Background images

Figure 5.4 Examples from the image database. (a) The foreground images consisted of 120 doll images that rotated over up to 180º angle of rotation; (b) the set of background images contained different objects, such as fighters, cars, bears and cups.

This indicates that some overfitting remained and that the amount of training data was not sufficient to estimate all degrees of freedom of the model. For a larger set of training images, we expect both errors to approach a common limit, somewhere between the observed test and training errors.

Table 5.1 compares the recognition results of the models trained over different rotation range. We first trained the model over 0º~60º rotation angle. One can observe that the models trained have narrow tuning characteristics. Also, these models are not superior to more broadly trained models within their designed rotation range. This suggests that a more diverse training set is generally beneficial. The models trained over 40º~100º show the best detection performance and good generalization. The likely reason for the somewhat disappointing performance of the model trained on the entire rotation range is that the learning algorithm is not able to identify consistent feature arrangements well in a very diverse data set.

(Training over) Rotation Range	0°~60°	40°~100°	60°~120°	0°~120°
Recognition Rate (%Correct)	88%	94%	81%	90%

Table 5.1: A comparison of the classification performance of models trained over different object rotation rages

5.6 Discussion and Future Work

We have demonstrated that our model works successfully in an unsupervised setting. Our model represents objects as probabilistic constellations of parts automatically extracted by the method described in chapter 4. The variability within a class is represented by a joint probability density function on the shape of the constellation and the appearance of the parts. The set of model parameters is then learned using EM. Experiments show good generalization performance to novel viewpoints. Performance is generally above 90% correct with only a few training examples.

Traditionally, object recognition starts with a training set where the salient parts of each object are in rough geometrical correspondence. This training set is either obtained by segmentation and warping/alignment of each object’s picture, or by direct manual identification of the main features. In either case, intervention of an operator and/or controlled imaging conditions are required (Burl, Leung et al. 1996; Cootes and Taylor 1996; Amit and Geman 1999). Amit and Geman have developed a method for visual selection which learns a hierarchical model starting with a simple type of feature detector (edge elements) at the front end. The method assumes that training images are registered with respect to a reference grid. After an exhaustive search through all possible local feature detectors, a global model is built, under which shape variability is encoded in the form of small regions in which local feature can move freely.

Burl (Burl, Weber et al. 1998) have proposed a statistical model in which shape variability is modeled in a probabilistic setting using Dryden-Mardia shape space densities (Burl, Leung et al. 1996). Their method requires labeled part positions in the training images. Similar approaches to object recognition include the active appearance models of Taylor et al. (Cootes and Taylor 1996) who model global deformations using Eigenspace methods as well as the Dynamic Link Architecture of v. der Malsburg and colleagues, who compute the “deformation energy” of a grid that links landmark points on the surface of objects (Lades, Vorbruggen et al. 1993). Yuille has also proposed a recognition method based on gradient descent on a deformation energy function in (Yuille 1991). It is not obvious how these methods would be trained without supervision.

Our model improves upon previous work on object recognition in two ways. It is viewpoint invariant, rather than restricted to certain fixed views of the object. Also, no direct supervision is required for training the system, unlike previous work where an operator had to align and normalize the training set and/or identify the most distinctive features/parts of each training example. Furthermore, our method is more biologically plausible by integrating unsupervised learning algorithms and sparse coding constraints into “ventral pathway” like hierarchical networks to automatically extract object parts for object recognition.

Moreover, the image sets used in our experiments are the same as those in Nelson’s appearance-based object recognition system (Nelson and Selinger 1998). In Nelson’s system, potential key features and local contexts of a target object are extracted by a stick growing algorithm (Nelson 1994). These features are used to access the database memory, retrieve match information, and generate hypotheses about the identity and configuration of the objects that could have produced them. Then, loosely consistent groupings of these “pose” hypotheses are identified. This integration is performed by using the pose hypotheses themselves as keys into a second associative memory, where evidence for the various global hypotheses is accumulated. If no match to a preexisting global hypothesis, a new one is generated. The performance of Nelson’s system was around 90%, when trained on a large set of training examples. Comparably, our model learned from just a few training examples can achieve same accurate recognition.

However, some aspects of our system are suboptimal and so there are a number of opportunities for improving system performance. Rather than classifying every image by applying a fixed decision threshold according to Eq. (5.14), we may compute receiver operating characteristics (ROCs) (Weber, Welling et al. 2000) based on the ratio of posterior probabilities. We determine the point on the ROC curve corresponding to an equal fraction of misclassified foreground and background images and use this error rate as a performance measure. On model training, our greedy algorithm is not necessarily optimal. It is likely that the performance of the system would further improve if more and more diverse training examples were used.

A number of issues of our object recognition model are still unexplored. First and foremost, more comprehensive experiments need to be carried out on a larger number of categories, in order to understand how categorical similarity affects the process. Second, in order to make our experiments practical we have simplified the probabilistic models that are used for representing objects. For example a probabilistic models that are used for occlusion is not implemented in our experiments. Third, it would be highly valuable for practical applications to develop an incremental version of our algorithm, where each training example will incrementally update the probability density function defined on the parameters of each object category. Fourth, the scale information of objects (or object parts) should also be incorporated into the model. In addition, the minimal training set and learning time that appear to be required by our algorithm makes it possible to conceive of visual learning applications where real-time training and user interaction are important.

Chapter 6

Conclusions and Future Work

One of the most fundamental properties of the primate visual system is the ability to recognize a particular object, despite great variations in the images that impose on the retina. While it may be obvious that primates are capable of recognizing objects under many variations in conditions, it is very difficult for object recognition in machine visions largely invariant with regard to changes in the size, position, and viewpoint of the object. Especially, invariance in viewpoint provides the greatest challenge to object recognition theory.

Two main computational theories have been proposed to address invariant object recognition: structural description theory and viewpoint-specific theory. However, structural description theories and view-specific theories have their own strengths and weaknesses. They are inadequate to capture the astonishing power, versatility, and subtlety of human object recognition. How people recognize objects is certainly among the most difficult problems in visual perception, so difficult that no satisfactory solution has yet been proposed for how the system actually works.

In this dissertation, we presented a solution to automatically extract parts (i.e. geons or primitive volumetric components) from raw gray-scale images, therefore, which addresses the most difficult problem of classic structural description theories: lack of representational power. However, when an object is rotated away from the training viewpoint, the part structural description of the object is altered accordingly. These changes may reduce generalization capability of structural description methods, thus causing problems for object recognition with rotational invariance. We therefore combined the view-specific theory with structural description theory by training models with multiple views of an object, which results in a more complete structural description, thereby allowing for recognizing object with rotational invariance.

Within a structural description framework, our computational scheme for invariant object recognition used the world as its own representation and extracted information from it through the action of feature detectors based on the notion of receptive fields (RFs). Simulated with the primate visual system, the scheme described in Chapter 2 is essentially composed of three modules: the visual ventral pathway module, the bottom-up saliency module and the top-down attention module. Using bottom-up, image-based saliency cues and top-down, task-dependent cues, visual information processed in the ventral pathway module focuses on what is important in a given situation, which reduces the computational complexity and the system memory.

As one moves through the primate ventral pathway that underlies object recognition, the RF size of cells increases. This property of cells indicates a hierarchical organization in the ventral pathway. Inspired by the ventral pathway, we developed a series of models in a hierarchical architecture by sparse and unsupervised learning. Object recognition was envisioned as a hierarchical process, with progressively more complex features at each stage. Local orientation of edges was considered to be the primary shape feature in early stages, based on the prevalence of orientation tuning in areas V1 and V2. Shape representation involved structural description based on parts and position relations of parts similar to that in area V4. The final object recognition was achieved in the inferior temporal (IT) model based on the output of the lower stages.

Followed previous work (Bell and Sejnowski 1997; Hyvarinen and Hoyer 2000), we developed the V1 model by extending independent component analysis (ICA) to combine the principle of invariant-feature subspaces and the multi-dimensional ICA. Similar to the RFs of cells in the V1 area, learned V1 model cells (M-cells) are oriented, localized in space and time, and displays phase invariance and limited shift invariance. Adding a layer on top of the V1 model, we developed a V2 model as a two-layer network (including V1 layer and V2 layer). Instead of computing the V1 complex M-cell response by a simple energy model as precious work (Hoyer 2002), our V2 model used the end-to-end learned RFs, and thus represented the natural image sparsely yet sufficiently. The active patterns of V2 M-cells are in different positions, different orientations and different lengths. The orientation and length tunings of our V2 M-cells are richer than what has been reported before (Rao and Ballard 1999; Hoyer 2002) and they also exhibit the end-stopped properties observed in real V2 cortex.

Shape representation is a major challenge in machine vision and is the basis for object recognition. Area V4 is an important intermediate stage of the ventral pathway and provides the major input to the final stages in the IT cortex for object recognition. Although originally thought to be primarily involved in color vision, V4 is now recognized to be critical for shape representation. In Chapter 4, we described a parts-based shape representation model – V4 model. Combining unsupervised representation in the feed-forward stream with lateral interaction, our V4 model can automatically decompose objects into parts and successfully achieve stable, efficient and natural description of shapes. For validating the reliability of learned parts, we used temporal Konhonen Map (TKM) algorithm to project parts learned in continuous time according to their mutual similarities. As the spin-off of the reliability analysis, we have demonstrated that our methods can successfully compress the representative parts of the same object over 180° rotation angle, which indicates that it is possible to achieve object recognition with rotation invariance by training the models with multiple views of an object. Comparing the tuning properties of V4 M-cells to those physiological measurements of V4 biological cells (Pasupathy and Connor 2001) via the "MAX" pooling mechanism, the experimental results show that our V4 model is able to reproduce the reported physiological findings. Similar to the V4 biological cells, the M-cells respond to a wide variety of shapes, but do not appear to encode any single type of a global shape. The shapes evoking strongest responses are characterized by a consistent type of boundary element at a specific position within the stimulus.

During the past decade, there have been major advances in our understanding of how object recognition is performed in the primate visual system. There is now a broad body of evidence (Saleem, Cheng et al. 1995; Tanaka 1996; Booth and Rolls 1998; Kobatake, Wang et al. 1998; Haxby, Gobbini et al. 2001; Rolls, Aggelopoulos et al. 2003) showing that object recognition makes use of cells in IT cortex that respond to features of intermediate complexity, such as object parts. To quantify the utility of the learned V4 M-cell RFs (i.e. learned object parts) for representing objects and for supporting various visual tasks, such as object recognition in IT, we presented a method to learn an IT model based on object parts learned in V4 area. Following previous work (Burl, Weber et al. 1998; Weber, Welling et al. 2000), we modeled objects as flexible constellations of learned parts. The variability within an object class is represented by a joint probability density function (pdf) on the shape of the constellation. The set of model parameters, including the shape pdf, is then learned using the Expectation Maximization (EM) algorithm.

Our IT model improved over previous object recognition models in three ways. First, our model is viewpoint/rotation invariant, rather than restricted to a fixed view of the object. Also, no direct supervision is required for training the system, unlike previous work where an operator had to align and normalize the training set and/or click on the most distinctive features/parts of each training example. Second, our model is more biologically plausible since it integrates an unsupervised learning algorithm and sparse coding constraints into the hierarchical network. And third, our model can learn object categories from just a few training images and still achieve good recognition performance.

In summary, the models developed here can be distinguished from previous work (Mel 1997; Nelson and Selinger 1998; Riesenhuber and Poggio 1999; Weber, Welling et al. 2000) by their use of three biologically plausible strategies: hierarchical organization, unsupervised learning and sparse coding.

Hierarchical organization draws inspiration from anatomical work showing that known or probable ascending connections originate in the lower cortical layers and project principally to layer 4 (Jones and Wise 1977) and within striate cortex one can conceive a hierarchy of neurons with increasingly complex RFs (Livingstone and Hubel 1987). Also, a hierarchical organization is beneficial because different modules need not duplicate the machinery required for shared components of the analysis. Our visual models described in this dissertation were hierarchically connected, with increasingly sophisticated representation from low to high layers.

Biological research (Edelman, Intrator et al. 2002) shows that unsupervised methods are essential for visual information processing in the primate brain, and thus, can be used as computational strategies for developing advanced visual modeling that is both highly sophisticated and versatile. In this dissertation, our models were developed by unsupervised learning based on statistical properties.

Biologically based principles, such as sparse coding, will likely have the information processing capabilities as well as huge payoffs in power/energy minimization and optimal resource management (Barlow 1994). Sparse coding has been proven to provide superior information storage capacity compared to local (grandmother cell theory or Gnostic representations) or fully distributed information representation (Attwell and Laughlin 2001). Because only a very few cells need to be activated and there are only a few cells encoding an event, sparse coding constraints used in our modeling work can bring fault-tolerance and low-power implementation (Jabri 2000) to the physical realization of computational models.

The main contribution of this work is the biologically motivated integration of a number of existing approaches, e.g., unsupervised learning and sparse representation, into the hierarchical network architecture. These models yield better performance than many existing algorithms and represent biologically plausible mechanisms, and therefore, may provide a platform from where to further explore the mechanisms of visual information processing both in biological and robotic settings.

In spite of its success, our object recognition system still needs improvement in training time and recognition accuracy for real applications. The bottom-up salience and top-down attention should also be better integrated to the system, to enable invariant object recognition on images with complex backgrounds. Area IT is the highest and most complicated pure visual area in the ventral pathway. We should explore more interesting properties of cells in IT.

The prefrontal cortex (PFC) constitutes the highest level of the cortical hierarchy dedicated to the representation and execution of actions. PFC is reciprocally connected with visual cortices, especially with IT. Several lines of evidence (Fuster, Bauer et al. 1985; Desimone 1996; Asaad, Rainer et al. 1998; Miller 1999; Asaad, Rainer et al. 2000; Freedman, Riesenhuber et al. 2001; Freedman, Riesenhuber et al. 2003) suggest that PFC is a major source of the proposed top-down inputs to the ventral pathway. Therefore, the PFC model would be necessary to complete the “ventral pathway”-like modeling work.

References

Amari, S. (1993). Neural representation of information by sparse encoding. Brain Mechanisms of Perception and Memory from Neuron to Behavior. T. Ono et. al, Oxford, UK, Oxford University Press**:** 630-637.

Amari, S., A. Cichocki, et al. (1996). A new learning algorithm for blind signal separation. Advances in Neural Information Processing Systems, D. Touretzky, Cambridge, MA, The MIT Press. **8**: 757-763.

Amit, Y. and D. Geman (1999). "A computational model for visual selection." Neural Computation **11**(7): 1691-1715.

Asaad, W. F., G. Rainer, et al. (1998). "Neural activity in the primate prefrontal cortex during associative learning." Neuron **21**(6): 1399-1407.

Asaad, W. F., G. Rainer, et al. (2000). "Task-specific neural activity in the primate prefrontal cortex." Journal of Neurophysiology **84**: 451-459.

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

Attwell, D. and S. Laughlin (2001). "An energy budget for signaling in the grey matter of the brain." Journal of Cerebral Blood Flow Metabolism **21**: 1133-1145.

Bandyopadhay, S. and U. Maulik (2001). "Nonparametric genetic clustering: Comparison of validity indices." IEEE Transactions on Systems, Man and Cybernetics C **31**(1): 120–125.

Barlow, H. (1989). "Unsupervised learning." Journal of Neural Computation **1**: 295-311.

Barlow, H. (1994). What is the computational goal of the Neocortex? Large-Scale Neuronal Theories of the Brain. C. Koch and J. Davis, Oxford, UK, Oxford University Press: 553-561.

Bartlett, M. S., H. M. Lades, et al. (1998). Independent component representations for face recognition. In Proceedings of the SPIE Conference on Human Vision and Electronic Imaging III: 528-539, San Jose, CA.

Bartlett, M. S. and T. J. Sejnowski (1997). Independent components of face images: A representation for face recognition. In Proceedings of the 4th Annual Joint Symposium on Neural Computation: 547-553, Pasadena, CA.

Bartram, D. J. (1974). "The role of visual and semantic codes in object naming." Cognitive Psychology, **6**: 325-356.

Baum, E. B., J. Moody, et al. (1988). "Internal representations for associative memory." Biological Cybernetics **59**: 217-228.

Becker, S. (1991). "Unsupervised learning procedures for neural networks." International Journal of Neural Systems **2**: 17-33.

Becker, S. and M. Plumbley (1996). "Unsupervised neural network learning procedures for feature extraction and classification." Applied Intelligence **6**: 185-203.

Beeck, H. O. D. and R. Vogels (2000). "Spatial sensitivity of Macaque inferior temporal neurons." The Journal of Comparative Neurology **426**: 505-518.

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

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

Bezdek, J. and N. Pal (1998). "Some new indexes of cluster validity." IEEE Transactions on Systems, Man and Cybernetics B **28**: 301-315.

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

Biederman, I. and M. Bar (1999). "One-shot viewpoint invariance in matching novel objects." Vision Research **39**: 2885-2899.

Biederman, I. and E. E. Cooper (1991). "Evidence for complete translational and reflectional invariance in visual object priming." Perception **20**: 585-593.

Biederman, I. and E. E. Cooper (1992). "Size invariance in visual object priming." Journal of Experimental Psychology: Human Perception and Performance **18**: 121-133.

Biederman, I. and P. C. Gerhardstein (1993). "Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance." Journal of Experimental Psychology: Human Perception and Performance **19**: 1162-1182.

Biederman, I. and P. C. Gerhardstein (1995). "Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995)." Journal of Experimental Psychology: Human Perception and Performance **21**: 1506-1514.

Binford, T. O. (1971). Visual perception by computer. In Proceedings of the IEEE Conference on Systems and Controls: 21-26, Miami, FL.

Booth, M. C. A. and E. T. Rolls (1998). "View-invariant representation of familiar objects by neurons in the inferior temporal visual cortex." Cerebral Cortex **8**(6): 510-523.

Boussaoud, D., R. Desimone, et al. (1991). "Visual topography of area TEO in the macaque." The Journal of Comparative Neurology **306**(4): 554-575.

Burkhalter, A., K. L. Bernardo, et al. (1993). "Development of local circuits in human visual cortex." Journal of Neuroscience **13**: 1916-1931.

Burl, M. C., T. K. Leung, et al. (1996). Recognition of planar object classes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96): 223-230, San Francisco, CA.

Burl, M. C., M. Weber, et al. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In Proceedings of the 5th European Conference on Computer Vision (ECCV'98): 628-641, Freiburg, Germany.

Buxton, H. and B. Neumann (1996). Visual interpretation and understanding, http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/ECVNET/BUXTON//interpretation.html, CVonline: Recognition and Registration Methods.

Cannon, M. W. and S. C. Fullenkamp (1991). "Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations." Vision Research **31**: 1985-1988.

Cardoso, J. F. (1998). Multidimensional independent component analysis. In Proceedings of the IEEE 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98) **4**: 1941-1944, Seattle, WA.

Chappell, G. and J. Taylor (1993). "The temporal Kohonen map" Neural Networks **6**: 441–445.

Cootes, T. F. and C. J. Taylor (1996). Locating objects of varying shape using statistical feature detectors. In Proceedings of the 4th European Conference on Computer Vision-Volume II: 465-474, Springer-Verlag, London, UK.

Daugman, J. G. (1985). "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters." Journal of the Optical Society of America A **2**(7): 1160-1169.

Dayan, P., G. E. Hinton, et al. (1995). "The Helmholtz machine." Journal of Neural Computation **7**(5): 889-904.

Deco, G. (2001). Biased competition mechanisms for visual attention in a multimodular neurodynamical system." Emergent Neural Computational Architectures Based on Neuroscience (LNAI 2036): 114-126, Springer-Verlag, Berlin Heidelberg.

Dempster, A. P., N. M. Laird, et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of Royal Statistical Society, Series B **39**(1): 1-38.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. In Proceedings of the National Academy of Sciences of the United States of America **93**(24): 13494-13499, Irvine, CA.

Desimone, R. and J. Duncan (1995). "Neural mechanisms of selective visual attention." Annual Review of Neuroscience **18**: 193-222.

Desimone, R. and S. Schein (1987). "Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form." Neurophysiology **57**(3): 835-868.

Desimone, R., S. J. Schein, et al. (1985). "Contour, color and shape analysis beyond the striate cortex." Vision Research **25**(3): 441-452.

Duncan, J. (1980). "The locus of interference in the perception of simultaneous stimuli." Psychological Review **87**: 272-300.

Edelman, S. (1997). "Computational Theories of Object Recognition." Trends in Cognitive Sciences **1**: 296--304.

Edelman, S. and H. H. Bülthoff (1992). "Orientation dependence in the recognition of familiar and novel views of three-dimensional objects." Vision Research **32**(12): 2385-2400.

Edelman, S., B. P. Hiles, et al. (2002). Probabilistic principles in unsupervised learning of visual structure: Human data and a model, Brain Computer Interface Workshop (NIPS*2001), Cambridge, MA, The MIT Press: 368-375.

Edelman, S., N. Intrator, et al. (2002). "Unsupervised learning of visual structure." Biologically Motivated Computer Vision 2002: 629-642.

Ellis, R., D. A. Allport, et al. (1989). "Varieties of object constancy." Quarterly Journal of Experimental Psychology **A41**: 775-796.

Fahle, M. (1997). "Specificity of learning curvature, orientation, and vernier discriminations." Vision Research **37**(14): 1885-1895.

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." Cerebral Cortex **1**: 1-47.

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

Fiser, J. and R. N. Aslin (2001). "Unsupervised statistical learning of higher-order spatial structures from visual scenes." Psychological Science **12**(6): 499-504.

Fiser, J. and I. Biederman (1995). "Size invariance in visual object priming of gray-scale images." Perception **24**: 741-748.

Foldiak, P. and M. Young (1995). Sparse coding in the primate cortex. The Handbook of Brain Theory and Neural Networks. Cambridge, MA, The MIT Press**:** 895-898.

Freedman, D. J., M. Riesenhuber, et al. (2001). "Categorical representation of visual stimuli in the primate prefrontal cortex." Science **291**: 312-316.

Freedman, D. J., M. Riesenhuber, et al. (2003). "A comparison of primate prefrontal and inferior temporal cortices during visual categorization." Journal of Neuroscience **23**(12): 5235-5246.

Fuster, J. M., R. H. Bauer, et al. (1985). "Functional interactions between inferotemporal and prefrontal cortex in a cognitive task." Brain Research **330**: 299-307.

Gallant, J. L., J. Braun, et al. (1993). "Selectivity for polar, hyperbolic, and Cartesian gratings in Macaque visual cortex." Science **259**: 100-103.

Gallant, J. L., C. E. Connor, et al. (1996). "Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey." Journal of Neurophysiology **76**(4): 2718-2739.

Gibson, J. J., G. A. Kaplan, et al. (1969). "The change from visible to invisible: A study of optical transition." Perception & Psychophysics **43**(5): 415-424.

Gibson, J. J. (1950). The perception of the visual world. In Proceedings of Computer Graph Image **9**:301-315, Boston, Houghton Mifflin.

Gilbert, C. D., J. A. Hirsch, et al. (1990). Lateral interactions in visual cortex. In Cold Spring Harbor Symposia on Quantitative Biology **55**:663-677, Cold Spring Harbor Laboratory Press.

Gilbert, C. D. and T. N. Wiesel (1979). "Morphology and intracortical projections of functionally identified neurons in cat visual cortex." Nature **280**: 120-125.

Ginsburg, A. P. (1971). Psychological correlates of a model of the human visual system. IEEE Transactions on Aerospace and Electronic Systems: 309-316.

Ginsburg, A. P. (1986). Spatial filtering and visual form perception. Handbook of Perception and Human Performance. K. R. Boff, L. Kauffman and J. P. Thomas. New York, John Wiley & Sons, Inc. **2:** 1-41.

Gottlieb, J. P., M. Kusunoki, et al. (1998). "The representation of visual salience in monkey parietal cortex." Nature **391**: 481-484.

Graham, N. (1989). Visual Pattern Analyzers. New York, Oxford University Press.

Hamker, F. H. and J. Worcester (2002). Object detection in natural scenes by feedback. Biologically Motivated Computer Vision Workshop (BMCV'02): 339-348, Tubingen, Germany, Springer-Verlag.

Hartline, H. K. (1938). "The response of single optic nerve fibers of the vertebrate eye to illumination of the retina." American Journal of Physiology **121**: 400-415.

Hateren, J. H. v. and A. v. d. Schaaf (1998). Independent component filters of natural compared with simple cells in primary visual cortex. In Proceedings of the Royal Society B: Biological Sciences **B, 265**: 359-366, Berlin Heidelberg, Springer-Verlag.

Haxby, J. V., M. I. Gobbini, et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex." Science **293**(5539): 2425-30.

Hayward, W. G. and M. J. Tarr (1997). "Testing conditions for viewpoint invariance in object recognition." Journal of Experimental Psychology: Human Perception and Performance **23**: 1511-1521.

Hebb, D. O. (1949). The Organization of Behavior. New York, Wiley.

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

Heeley, D. W. and H. M. Buchanan-Smith (1996). "Mechanisms specialized for the perception of image geometry." Vision Research **36**(22): 3607-3627.

Himberg, J., A. Hyvärinen, et al. (2004). "Validating the independent components of neuroimaging time-series via clustering and visualization." NeuroImage **22**(3): 1214-1222.

Hinton, G. and T. Sejnowski (1986). Learning and relearning in Boltzmann machines. Parallel Distributed Processing: Explorations in Microstructure of Cognition. D. E. Rumelhart, J. L. McClelland and t. P. r. group. Cambridge, MA, The MIT Press. **1:** 282-317.

Hinton, G. and T. J. Sejnowski, Eds. (1999). Unsupervised Learning: Foundations of Neural Computation. Cambridge, MA, The MIT Press.

Hinton, G. and R. Zemel (1994). Autoencoders, minimum description length, and Helmholtz free energy. Advanced in Neural Information Processing Systems. J. Cowan, G. Tesauro and J. Alspector. San Francisco, CA, Morgan Kaufmann. **6:** 3-10.

Hinton, G. E., P. Dayan, et al. (1995). "The wake-sleep algorithm for unsupervised neural networks." Science **268**: 1158-1161.

Hoffman, D. D. and W. A. Richards (1984). "Parts of recognition." Cognition **18**: 65-96.

Holmes, G. (1918). "Disturbances of vision by cerebral lesions." British Journal of Ophthalmology **2**: 353-384.

Hoyer, P. O. (2002). "Modeling receptive fields with non-negative sparse coding." Neurocomputing **52**(54): 547-552.

Hoyer, P. O. and A. Hyvarinen (2002). "A multi-layer sparse coding network learns contour coding from natural images." Vision Research **42**(12): 1593-1605.

Hubel, D. H. (1963). "The visual cortex of the brain." Scientific American **209**(5): 54-62.

Hubel, D. H. (1995). Eye, Brain, and Vision, Scientific American Library.

Hummel, J. E. (1999). The MIT Encyclopedia of the Cognitive Sciences on-line: Binding Problem & Binding by Neural Synchrony, CogNet library. **2002**. http://cognet.mit.edu/MITECS/Articles/hummel.html. Date viewed: Aug. 21, 2003.

Hummel, J. E. (2000). Where view-based theories of human object recognition break down: the role of structure in human shape perception. Cognitive Dynamics: conceptual change in humans and machines. E. Dietrich and A. Markman. Hillsdale, NJ, Erlbaum: 157-185.

Hummel, J. E. and I. Biederman (1992). "Dynamic binding in a neural network for shape recognition." Psychological Review **99**: 480-517.

Huttenlocher, D. P. and S. Ullman (1987). Object recognition using alignment. MIT-AI Memo 937, Cambridge, MA, The MIT Press.

Hyvarinen, A. and P. Hoyer (2000). "Emergence of Topography and Complex Cell Properties from Natural Images using Extensions of ICA." Advances in Neural Information Processing Systems **12 (**NIPS'00**)**: 827-833.

Hyvarinen, A. and P. O. Hoyer (2000). "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces." Neural Computation **12**(7): 1705-1720.

Intrator, N. (1992). "Feature extraction using an unsupervised neural network." Neural Computation **4**: 98-107.

Ito, M., H. Tamura, et al. (1994). "Size and position invariance of neuronal responses in monkey inferotemporal cortex." Journal of Neurophysiology **73**(1): 218-226.

Itti, L. and C. Koch (2000). "A saliency-based search mechanism for overt and covert shifts of visual attention." Vision Research **40**(10-12): 1489-1506.

Itti, L. and C. Koch (2001). "Computational modeling of visual attention." Nature Reviews Neuroscience **2**(3): 194-203.

Itti, L. and C. Koch (2001). "Feature combination strategies for saliency-based visual attention systems,." Journal of Electronic Imaging **10**(1): 161-169.

Itti, L., C. Koch, et al. (1998). "A model of saliency-based visual attention for rapid scene analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **20**: 1254-1259.

Jabri, M. (2000). Biological computing for robot navigation and control. NRA2-37143. Beaverton, OR, OGI School of Science and Engineering at OHSU.

Jones, E. G. and S. P. Wise (1977). "Size, laminar and columnar distribution of efferent cells in the sensory-motor cortex of monkeys." Journal of Comparative Neurology **175**: 391-437.

Kanwisher, N. and E. Wojciulik (2000). "Visual Attention: Insights from Brain Imaging." Nature Reviews Neuroscience **1**: 91-100.

Kirkpatrick, K. (2001). Object Recognition. Avian Visual Cognition. R. G. Cook, Berlin Heidelberg, Springer-Verlag.

Kobatake, E. and K. Tanaka (1994). "Neuronal selectivities to complex object features in the ventral visual pathway of the Macaque cerebral cortex." Journal of Neurophysiology **71**(3): 856-867.

Kobatake, E., G. Wang, et al. (1998). "Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys." Journal of Neurophysiology **80**: 324-330.

Koenderink, J. J. and A. J. van Doorn (1979). "The internal representation of solid shape with respect to vision." Biological Cybernetics **32**: 211-216.

Kohonen, T. (1995). Self-Organizing Maps. Berlin Heidelberg, Springer-Verlag.

Kohonen, T. (1996). "Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map." Biological Cybernetics **75**: 281-291.

Kohonen, T. and R. Hari (1999). "Where the abstract feature maps of the brain might come from." Trends in Neurosciences **22**: 135-139.

Lades, M., J. C. Vorbruggen, et al. (1993). "Distortion invariant object recognition in the dynamic link architecture." IEEE Transactions on Computations **42**(3): 300-311.

Larsson, J., K. Amunts, et al. (2002). "Perceptual segregation of overlapping shapes activates posterior extrastriate visual cortex in man." Experimental Brain Research **143**(1): 1-10.

LeCun, Y. and Y. Bengio (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. M. A. Arbib, Cambridge, MA, The MIT Press**:** 255-258.

Lee, D. D. and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." Nature **401**(6755): 788-791.

Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13 (NIPS'00), Cambridge, MA, The MIT Press: 556-562.

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

Lennie, P. (1998). "Single units and visual cortical organization." Perception **27**: 889-935.

Lennie, P. (2003). "The cost of cortical computation." Current Biology **13**: 493-497.

Leung, T. K., M. C. Burl, et al. (1998). Probabilistic affine invariants for recognition. In Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98): 678-684, Santa Barbara, CA.

Levine, E. and E. Domany (2001). "Resampling method for unsupervised estimation of cluster validity." Neural Computation **13**(11): 2573–2593.

Levy, W. and R. Baxter (1996). "Energy efficient neural codes." Neural Computation **8**: 531-543.

Li, S. Z., X. Hou, et al. (2001). Learning spatially localized, parts-based representation. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'01): 207-211, Hawaii.

Livingstone, M. S. and D. H. Hubel (1987). "Psychophysical evidence for separate channels for the perception of form, color, movement, and depth." Journal of Neuroscience **7**: 3416-3468.

Logothetis, N. K., J. Pauls, et al. (1995). "Shape representation in the inferior temporal cortex of monkeys." Current Biology **5**(5): 552-563.

Logothetis, N. K. and D. L. Sheinberg (1996). "Visual object recognition." Annual Review of Neuroscience **19**: 577-621.

Lowe, D. G. (2000). Towards a computational model for object recognition in IT cortex. In Proceedings of the First IEEE International Workshop on Biologically Motivated Computer Vision: 20-31, Seoul, Korea.

Löwel, S. and W. Singer (1992). "Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity." Science **255**: 209-212.

Luck, S. J., L. Chelazzi, et al. (1997). "Neural mechanisms of spatial selective attention in areas V1, V2 and V4 of Macaque visual cortex." Neurophysiology **77**: 24-42.

MacKay, D. (1956). The epistemological problem for automata. Automata Studies. C. E. Shannon and J. MeCarthy. Princeton, NJ, Princeton University Press**:** 235-251.

Marr, D. (1982). Vision. New York, W. H. Freeman and Company.

Marr, D. and H. K. Nishihara (1978). Representation and recognition of the spatial organization of three-dimensional shapes. In Proceedings of the Royal Society of London B Biology Science **200** (1140): 269-294, Berlin Heidelberg, Springer-Verlag.

Maulik, U. and S. Bandyopadhay (2002). "Performance evaluation of some clustering algorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **24**(12): 1650–1654.

Mel, B. W. (1997). "SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition." Neural Computation **9**: 777-804.

Merigan, W. H. (1996). "Basic visual capacities and shape discrimination after lesions of extrastriate area V4 in macaques." Visual Neuroscience **13**: 51-60.

Miller, E. K. (1999). "Straight from the top." Nature **401**(6754): 650-651.

Moran, J. and R. Desimone (1985). "Selective attention gates visual processing in the extrastriate cortex." Science **229**: 782-784.

Narendra, K. S. and K. Parthasarathy (1990). "Identification and control of dynamical systems using neural networks." IEEE Transactions on Neural Networks **1**: 4-27.

Nelson, R. C. (1994). "Finding line segments by stick growing." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **16**(5): 519-523.

Nelson, R. C. (2003). Boundary Extraction by Lineal Feature Growing. **2003** http://www.cs.rochester.edu/users/faculty/nelson/research/boundaries/boundaries.html. Date viewed: Oct. 25, 2004.

Nelson, R. C. and A. Selinger (1998). A cubist approach to object recognition. In Proceedings of International Conference on Computer Vision (ICCV'98): 614-621, Bombay, India.

Nevatia, R. (1982). Machine Perception, Prentice Hall, Englewood Cliffs, NJ.

Niebur, E. and C. Koch (1998). Computational architectures for attention. The attentive brain. R. Parasuraman. Cambridge, MA, The MIT Press**:** 163-186.

Olshausen, B. A., C. H. Anderson, et al. (1993). "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information." Neuroscience **13**: 4700-4719.

Olshausen, B. A. and D. J. Field (1997). "Sparse coding with an overcomplete basis set: A strategy employed by V1?" Vision Research **37**: 3311-3325.

Olshausen, B. A. and D. J. Field (2004). "Sparse coding of sensory inputs." Current Opinion in Neurobiology **14**: 481-487.

Oppenheim, A. and J. Lim (1981). "The importance of phase in signals." Proceedings of the IEEE **69** (5): 529-541.

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

Palmer, S. E. (1977). "Hierarchical structure in perceptual representation." Cognitive Psychology **9**(4): 441-474.

Palmer, S. E. (1980). "What makes triangles point: Local and global effects in configurations of ambiguous triangles." Cognitive Psychology **12**(3): 283-305.

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. Human and Machine Vision. J. Beck, B. Hope and A. Rosenfeld. New York, Academic Press**:** 269-339.

Palmer, S. E. (1989). Reference frames in the perception of shape and orientation. Object Perception: Structure and Process. B. E. Shepp and S. Ballesteros. Hillsdale, NJ, Erlbaum: 127-139.

Palmer, S. E. (1999). Vision Science-Photons to Phenomenology. Cambridge, MA, The MIT Press.

Palmer, S. E. and N. M. Bucher (1981). "Configural effects in perceived pointing of ambiguous triangles." Journal of Experimental Psychology: Human Perception & Performance **8**(5): 693-708.

Palmer, S. E., E. Rosch, et al. (1981). Canonical perspective and the perception of objects. Attention and Performance. J. Long and A. Baddeley. Erlbaum, NJ, Hillsdale **9:** 135-151.

Pasupathy, A. and C. E. Connor (1999). "Responses to contour features in Macaque area V4." Journal of Neurophysiology **82**: 2490-2502.

Pasupathy, A. and C. E. Connor (2001). "Shape representation in area V4: position-specific tuning for boundary conformation." Journal of Neurophysiology **86**: 2505-2519.

Pasupathy, A. and C. E. Connor (2002). "Population coding of shape in area V4." Nature Neuroscience **5**(12): 1332-1338.

Pentland, A. P. (1987). Recognition by parts. In Proceedings of the IEEE International Conference on Computer Vision: 612-620, London, England.

Phillips, G. C., S. M. Zeki, et al. (1984). "Localization of function in the cerebral cortex: past, present and future." Brain Research **107**: 327-361.

Piotrowski, L. and F. Campbell (1982). "A demonstration of the visual importance and flexibility of spatial-frequency, amplitude, and phase." Perception **11**(3): 337-346.

Polat, U. and D. Sagi (1993). "Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments." Vision Research **33**: 993-999.

Polat, U. and D. Sagi (1994). "Spatial interactions in human vision: From near to far via experience-dependent cascades of connections." In Proceedings of the National Academy of Sciences **91**: 1206-1209.

Pomerantz, J. R., L. C. Sager, et al. (1977). "Perception of wholes and of their component parts: Some configural superiority effects." Journal of Experimental Psychology: Human Perception & Performance **3**(3): 422-435.

Rao, K. (1988). Shape description from sparse and imperfect data, IRIS Technical Report 250, PhD Thesis, University of Southern California.

Rao, R. P. N. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." Nature Neuroscience **2**(1): 79-87.

Reicher, G. M. (1969). "Perceptual recognition as a function of meaning fullness of stimulus material." Journal of Experimental Psychology **81**(2): 275-280.

Reynolds, J. H., L. Chelazzi, et al. (1999). "Competitive mechanisms subserve attention in Macaque areas V2 and V4." The Journal of Neuroscience **19**: 1736-1753.

Reynolds, J. H. and R. Desimone (1999). "The role of neural mechanisms of attention in solving the binding problem." Neuron **24**: 19-29.

Reynolds, J. H. and R. Desimone (2003). "Interacting roles of attention and visual salience in V4." Neuron **37**: 853-863.

Reynolds, J. H., T. Pasternak, et al. (2000). "Attention increases sensitivity of V4 neurons." Neuron **26**: 703-714.

Riesenhuber, M. and T. Poggio (1999). "Hierarchical models of object recognition in cortex." Nature Neuroscience **2**: 1019-1025.

Rock, I. and J. DiVita (1987). "A case of viewer-centered object perception." Cognitive Psychology **19**(2): 280-293.

Rolls, E. T., N. C. Aggelopoulos, et al. (2003). "The receptive field of inferior temporal cortex neurons in natural scenes." The Journal of Neuroscience **23**(1): 339-348.

Saleem, K. S., K. Cheng, et al. (1995). "Anatomical organization of afferent and efferent connections of the inferotemporal cortex." RIKEN Review No. 9: Focused on Brain and Information Science **9**: 5-6.

Schiller, P. H. (1995). "Effect of lesion in visual cortical area V4 on the recognition of transformed objects." Nature **376**(27): 342-344.

Schneiderman, H. and T. Kanade (2000). "A statistical approach to 3D object detection applied to faces and cars." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00): 746-751.

Schwark, H. D. and E. G. Jones (1989). "The distribution of intrinsic cortical axons in area 3b of cat primary somatosensory cortex." Experimental Brain Research **78**: 501-513.

Sejnowski, T. J. and C. R. Rosenberg (1987). "Parallel networks that learn to pronounce english text." Complex System **1**: 145-168.

Sereno, A. and J. Maunsell (1998). "Shape selectivity in primate lateral intraparietal cortex." Nature **395**: 500-503.

Shams, L. and C. v. d. Malsburg (2002). "Acquisition of visual shape primitives." Vision Research **42**: 2105-2122.

Shepard, R. N. and J. Metzler (1972). "Mental rotation of three-dimensional objects." Science **171**: 701-703.

Simoncelli, E. P. (2003). "Vision and the statistics of the visual environment." Current Opinion in Neurobiology **13**: 144-149.

Simula, O. and J. Kangas (1995). Process monitoring and visualization using self-organizing maps. Neural Networks for Chemical Engineers. A. B. Bulsari. Amsterdam, Elsevier Science. **6:** 371-384.

Smith, A. T., K. D. Singh1, et al. (2001). "Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex." Cerebral Cortex **11**(12): 1182-90.

Subirana-Vilanova, J. B. and W. Richards (1996). "Attentional frames, frame curves and figural boundaries: the inside/outside dilemma." Vision Research **36**(10): 1493-1501.

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

Tanaka, K. (1996). "Inferotemporal cortex and object vision." Annual Review of Neuroscience **19**: 109-139.

Tanaka, K., H. Saito, et al. (1991). "Coding of visual images of objects in the inferotemporal cortex of the macaque monkey." Journal of Neurophysiology **66**: 170-189.

Tarr, M. J. and H. H. Bulthoff (1998). "Image-based object recognition in man, monkey, and machine." Cognition **67**: 1-20.

Tarr, M. J. and H. H. Bülthoff (1995). "Is human object recognition better described by geon structural descriptions or multiple views? Comment on Biederman and Gerhardstein (1993)." Journal of Experimental Psychology: Human Perception and Performance **21**: 1494-1505.

Tarr, M. J., P. Williams, et al. (1998). "Three-dimensional object recognition is viewpoint dependent." Nature Neuroscience **1**: 275-277.

Tovee, M. J., E. T. Rolls, et al. (1996). "Rapid visual learning in neurons of the primate temporal visual cortex." NeuroReprot **7**: 2757-2760.

Treisman, A. M. (1986). Properties, parts, and objects. Handbook of Perception and Human Performance. K. R. Boff, L. R. Kaufman and J. P. Thomas. New York, JohnWiley. **35:** 1-70.

Ullman, S. (1996). High-level vision: object recognition and visual cognition. Cambridge, MA, The MIT Press.

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

Ullman, S., M. Vidal-Naquet, et al. (2002). "Visual features of intermediate complexity and their use in classification." Nature Neuroscience **5**(7): 682-687.

Ultsch, A. and H. P. Siemon (1990). Kohonen's self organizing feature maps for exploratory data analysis. In Proceedings of the International Neural Network Conference (INNC'90): 305-308, Dordrecht, Netherlands.

Ungerleider, L. G. and M. Mishkin (1982). Two cortical visual systems. Analysis of visual behavior. D. G. Ingle, M. A. Goodale and R. J. Q. Mansfield. Cambridge, MA, The MIT Press: 515-543.

Ungerleider, L. G. and M. Mishkin (1982). Two cortical visual systems: Separation of appearance and location of objects. Analysis of visual behavior. D. L. Ingle, M. A. Goodale and R. J. W. Mansfield. Cambridge, MA, The MIT Press**:** 549-586.

Unser, M. (1994). "Fast Gabor-like windowed Fourier and continuous wavelet transforms." IEEE Signal Processing Letters **1**(5): 76-79.

Versavel, M., G. A. Orban, et al. (1990). "Responses of visual cortical neurons to curved stimuli and chevrons." Vision Research **30**: 235-248.

Vetter, T., T. Poggio, et al. (1994). "The importance of symmetry and virtual views in three-dimensional object recognition." Current Biology **4**: 18-23.

Viola, P. and M. Jones (2001). "Rapid object detection using a boosted cascade of simple features." In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01) **1**: 511-518.

Walker, K. N., T. F. Cootes, et al. (1999). Recognizing surfaces using three-dimensional textons. In Proceedings of the 7th IEEE International Conference on Computer Vision: 1010-1017, Washington, DC.

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

Weber, M., M. Welling, et al. (2000). Towards automatic discovery of object categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00): 101-109, Washington, DC.

Weber, M., M. Welling, et al. (2000). Unsupervised learning of models for recognition. In Proceedings of the 6th European Conference on Computer Vision (ECCV'00): 18-32, Dublin, Ireland.

Weisstein, N. and C. S. Harris (1974). "Visual detection of line segments: An object-superiority effect." Science **186**(4165): 752-755.

Wickerhauser, M. V. (1994). Adapted wavelet analysis from theory to software. Wellesley, MA, A K Peters, Ltd.

Wilkinson, F., T. W. James, et al. (2000). "An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings." Current Biology **10**: 1455-1458.

Wilson, H. R., F. Wilkinson, et al. (1997). "Concentric orientation summation in human form vision." Vision Research **37**: 2325-2330.

Wilson, M. E. and B. G. Cragg (1967). "Projections from the lateral geniculate nucleus in the cat and monkey." Journal of Anatomy **101**: 677-692.

Winger, E. P. (1960). "The unreasonable effectiveness of mathematics in the natural sciences." Communications On Pure & Applied Mathematics **XIII**: 1-14.

Wolfe, J. M. and S. C. Bennett (1997). "Preattentive object files: Shapeless bundles of basic features." Vision Research **37**(1): 25-43.

Yang, L. and M. Jabri (2003). Sparse visual models for biologically inspired sensorimotor control. In Proceedings of the Third International Workshop on Epigenetic Robotics: 131-138, Boston, MA.

Yang, L. and M. Jabri (2006). "Part decomposition by sparse and unsupervised learning." Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).

Yuille, A. L. (1991). "Deformable templates for face recognition." Journal of Cognitive Neuroscience **3**(1): 59-70.

Zeki, S. and A. Bartels (1998). "The autonomy of the visual systems and the modularity of conscious vision." Philosophical Transactions of the Royal Society B: Biological Sciences, **Lond. B 353**: 1911-1914.

Zeki, S. M. (1975). "Colour coding in rhesus monkey prestriate cortex." Brain Research **53**: 422-472.

Zeki, S. M. (1983). "Colour coding in the cerebral cortex: the responses of wave-length selective and colour-coded cells in monkey visual cortex to changes in wavelength composition." Neuroscience **9**: 767-781.