SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

DISSERTATION

John G. Keller
Captain, USAF

AFIT/DS/ENG/99-03

19990616 028

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DTIC QUALITY INSPECTED 4

AFIT/DS/ENG/99-03

SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

DISSERTATION

John G. Keller
Captain, USAF

AFIT/DS/ENG/99-03

Approved for public release; distribution unlimited

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the Department of Defense or the United States Government.

AFIT/DS/ENG/99-03

SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

DISSERTATION

Presented to the Faculty of the Graduate School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

John G. Keller, B.S.E.E., M.S.E.E., M.I.S.
Captain, USAF

June, 1999

Approved for public release; distribution unlimited

AFIT/DS/ENG/99-03

SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

John G. Keller, B.S.E.E., M.S.E.E., M.I.S.

Captain, USAF

Approved:

15 MAY 99
Dr Steven K. Rogers, Chairman

15 MAY 99
Dr Matthew Kabrisky

May 11, 1999
Dr Martin P. DeSimio

11 May 99
Dr Mark E. Oxley

17 May 99
Dr Paul I. King, Dean's Representative

Dr Robert A. Calico, Jr.
Dean, Graduate School of Engineering

Table of Contents

Page

List of Figures vi

List of Tables x

Abstract xi

I. Introduction 1
  1.1 Background 1
  1.2 Problem Statement and Scope 2
    1.2.1 Problem Statement 2
    1.2.2 Scope 2
    1.2.3 Research Contributions 3
  1.3 Dissertation Organization 3

II. The Human Visual System 5
  2.1 Introduction 5
  2.2 The Physiology of Primate Vision 5
    2.2.1 Early Visual Processing 5
    2.2.2 Primary Visual Cortex 7
    2.2.3 Other Spatial Visual Attributes 11
  2.3 Saccadic Behavior 13
  2.4 Summary 15

III. Why a Vision Model Motivated by Behavior? 16
  3.1 Introduction 16
  3.2 Rybak's Saccadic Vision Model 17
    3.2.1 Top Level Overview 17
    3.2.2 Attention Window (AW) 19

Page

3.2.3 Feature Detection 20

3.2.4 Invariant Transformation 22

3.2.5 Selecting the Next Fixation Point 23

3.2.6 Results and Limitations 24

3.3 Summary 26

IV. SCAN-IT: A New Model of Computational Vision 27

4.1 Introduction 27

4.2 SCAN-IT Overview 27

4.3 SCAN-IT's Visual Field Representation 28

4.3.1 The Gabor Filter for Oriented Bar/Edge Detection . 28

4.3.2 Constructing the Visual Field of View (FOV) 30

4.3.3 View-Based Standardization 32

4.3.4 Matrix Representation of the Field of View 34

4.4 Saccadic Attraction 35

4.5 Memorization 36

4.6 Search/Recognition 36

4.7 Summary 39

V. Experimental Results using SCAN-IT 40

5.1 Introduction 40

5.2 Experimental Setup and Results 41

5.2.1 Trial Set 1: Fundamental Capabilities 43

5.2.2 Trial Set 2: Word Recognition in Sentences 46

5.2.3 Trial Set 3: Memorization/Recognition of Faces in Complex Backgrounds 48

5.2.4 Trial Set 4: Memorization/Recognition of Faces in Complex Backgrounds Containing other Faces . . . 49

Page

5.2.5 Trial Set 5: Memorization/Recognition of cars in Complex Backgrounds Containing other Cars 53
5.2.6 Trial Set 6: Memorization/Recognition of Text and/or Cars in Target Images Comprised of Text and Cars . 55
5.2.7 Summary of Experimental Results 61
5.3 Using Context to Solve the Sensitivity-vs-Specificity Problem 61

VI. Video Retrieval: A Real-world SCAN-IT Demonstration 71
6.1 Introduction to Image Retrieval 71
6.2 The Problem: Finding Specified Video Imagery 72
6.3 Conclusion 75

VII. Conclusion 76
7.1 Research Contributions 76
7.2 Future Directions 76

Appendix A. Experimental Trial Data 79
A.1 Trial Set 1 Data 79
A.2 Trial Set 2 Data 80
A.3 Trial Set 3 Data 81
A.4 Trial Set 4 Data 82
A.5 Trial Set 5 Data 84
A.6 Trial Set 6 Data 87

Appendix B. Video Mosaic Processing 90
B.1 Introduction 90
B.2 Frame Registration and Fusion 91
B.3 Conclusion 95

Bibliography 96

Vita 101

List of Figures

Figure Page

1. Simplified Diagram of the Human Visual Tract [70] 6

2. a) Human Eye. b) Density of photoreceptors in the retina. The cone receptors are concentrated in the fovea, accounting for greater resolution of parts of the visual scene projected on the fovea [77]. . 7

3. Cortical Magnification in V1. The bottom image shows a slice through area V1 of a monkey, where the dark cells are those that were responding as the animal viewed the pattern in the upper image [70]. . . . 8

4. Typical contrast sensitivity function characterizing a photopic foveal neuron's response to harmonic stimuli [48, 77]. 9

5. Center-Surround Organization of a neural receptive field in the visual tract [77]. 10

6. Non-circular neural receptive fields found in V1. 10

7. Model proposed by Sekuler to explain motion specificity of neurons [70]. 11

8. a) Photograph of a girl b) Recording of eye movements during examination of the photograph [81] 13

9. Block Diagram of the Rybak Visual Perception Model [66] 18

10. Multiple Resolution Annuli Surrounding Foveation Point 20

11. Relative Coordinate System with Respect to Basic Edge [66]. 23

12. Two dimensional Sine and Cosine Gabor functions. (a,c) Spatial domain response (b,d) Fourier transform of Gabor function. 29

13. Block diagram showing the construction of the Field of View. . . . 30

14. Example of the construction of the edge intensity map. 31

15. Transformation of FOV into polar space. 33

16. Polar transformation example. 34

17. Vision Model in Memorization Mode. 37

18. Vision Model in Search/Recognition Mode. 38

Figure Page

19. a) One of the words to be memorized, b) a rotational variant of the word, and c) a scale and rotational variant of the word. d) A rotational variant of a non-memorized word. 44

20. a) Saccadic path obtained during object memorization. b) Saccadic path taken during successful recognition of target variant. c) Search path for the memorized object in a target image not containing the target 45

21. The three sentences serving as the base target images for the second set of trials. 46

22. a) Memorization scan of the word **falls**. b) Successful recognition (dashed line indicates the search path performed by SCAN-IT). c) Search path for the word **falls** in a sentence not containing the word 47

23. a) and b) Two of the images from which the faces for Trial Set 3 were extracted. c) One of the extracted faces to be memorized. 48

24. a) Memorization scan of one of the faces used for Trial Set 3. b) Successful recognition of the memorized face. c) Search of a target image not containing the memorized face. 48

25. The original Gatlin image and two of the six sub-images used for Trial Set 4 (from the Matlab© development environment). 51

26. a) Memorization scan of one of the faces used for Trial Set 3. b) Successful recognition of the memorized face. c) Search of a target image not containing the memorized face. 52

27. a) Two of the individual cars and two of the car mosaics used for Trial Set 5. 53

28. a) Memorization scan of one of the cars used for Trial Set 5. b) Successful recognition of the memorized car. c) Search of a target mosaic not containing the memorized car. 54

29. Two of the target images used for Trial Set 6. 55

30. Portions of a car examined during the memorization process after a) 1, b) 3, and c) 8 saccades. 56

Figure Page

31. a) and d) Memorization scan of one of the objects used for Trial Set 6. b) and e) Successful recognition of the memorized object. c) and f) Search of a target image not containing the memorized object. . 57

32. False Positive results from Trial Set 6. The top row shows the search path and false recognitions, and the bottom row shows the original memorization of the object after 8 foveations. 58

33. False Negative result from Trial Set 6. a) shows the memorization of the car, and b) shows the search scan path. 60

34. a) Discriminant boundary constructed to separate two classes of data to some precision determined by the designer. b) Example of data points perfectly separated by the boundary. c) Example of data points producing False Positives. 62

35. Plots of Minkowski distances calculated for each recognition instance reported by SCAN-IT. Each row corresponds to one of the four text objects, for which two object-absent and two object-present cases were tested (the left and right plots respectively). 66

36. Graphical example of calculation of $\hat{D}_{max}$, $\hat{D}_{min}$, and $\Delta\hat{D}$ for context-based recognition 67

37. Comparison of $\Delta\hat{D}$ for the eight object-present cases (solid line) and the eight object-absent cases (dashed line). 68

38. Frames from the video sequence used for demonstrating SCAN-IT. a) The first frame in the sequence. b) and c) Two of the intermediate frames. d) The last frame of the sequence. 73

39. a) Memorization of the face and last name of the reporter, each via a sequence of six scans. b) Subsequent recognition of the face and last name in a different frame. 74

40. Correct recognition of the face object in Frames 40 and 87. 75

41. The five objects memorized for Trial Set 1. These objects also served as target images. 79

42. Examples of the target image variants used for Trial Set 1. 79

Figure Page

43. The original three target images (non-rotated/scaled) tested against for Trial Set 2 (top row) and examples of three of the variants (bottom row). 80

44. The five face objects memorized for Trial Set 3. 81

45. The original five target images tested against for Trial Set 3 (top two rows) and examples of three of the variants (bottom row). 81

46. The six face objects memorized for Trial Set 4. 82

47. The original four target images tested against for Trial Set 4 (top two rows) and examples of two of the variants (bottom row). 83

48. The five car objects memorized for Trial Set 5. 84

49. The original 10 target images tested against for Trial Set 5. 85

50. Three examples of the variants used for Trial Set 5 testing. 86

51. The four text objects memorized for Trial Set 6. The same car objects used for Trial set 5 were also memorized here. 87

52. The original 10 target images tested against for Trial Set 6. 88

53. Three examples of the variants used for Trial Set 6 testing. 89

54. a,b) Two of the video frames used for this registration and fusion demonstration. c) Scanpath of the vision model during memorization of the first frame. d) Scanpath of the vision model during search of the subsequent frame. The dashed line indicates the saccadic search, while the solid line denotes discovery of the memorized path. 93

55. Registration and fusion of six frames of video data. 94

List of Tables

Table Page

1. Experimental goals and conditions 42
2. Trial 1 Results 45
3. Trial 2 Results 47
4. Trial 3 Results 49
5. Trial 4 Results 52
6. Trial 5 Results 54
7. Trial 6 Results 57
8. Trial 6 Fusion Results 58
9. Summary of Experimental Results 61

AFIT/DS/ENG/99-03

Abstract

This dissertation details the development of a new computational vision model motivated by physiological and behavioral aspects of the human visual system. Using this model, intensity features within an artificial visual field of view are extracted and transformed into a simulated cortical representation, and a saccadic guidance system scans this field of view over an object within an image to 'memorize' that object. The object representation is thus stored as a sequence of feature matrices describing sub-regions of the object. A new image can then be searched for the object (possibly scaled and rotated), where evidence of its presence is accumulated by finding sub-regions in the new image similar to those stored and with the same relative spatial configuration. A set of over 450 experimental trials demonstrates the model is capable of memorizing and then recognizing arbitrary objects within arbitrary images, as well as correctly rejecting images that do *not* contain the memorized object. A new context-based recognition paradigm is introduced that solves the problem of *a priori* assignation of recognition threshholds, and also can be generalized to solve threshholding problems commonly found in pattern recognition environments. A demonstration is provided of the model's applicability to real-world problems by memorizing a face and text string, and then successfully searching a video sequence for their presence.

SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

I. Introduction

1.1 Background

Man has long attempted to imbue into machines human characteristics and capabilities. From tasks as simple as differentiating between dark and light to the perhaps intractable construction of fully conscious, thinking computers, we attempt to automate what to us comes completely naturally and easily. With proven working models (our own bodies), it seems rational to believe that if we understand precisely how our various biological systems function, we may be able to design automated implementations to emulate those functions. Indeed researchers have made great strides into understanding how we function physiologically and neurologically, but large gaps in our knowledge still exist. Though the physiology of the human brain has been extensively mapped, there is still very little concensus as to how this tremendously complex system of firing neurons is able to work in concert to produce the living, breathing, functioning organisms we are.

Perceptual processing (e.g., seeing, hearing, feeling) is an excellent example of the above dichotomy: though major portions of the neural pathways contributing to the perceptual senses have been been mapped (that is, specific areas of the brain have been associated with specific sensory functions), it is still not at all clear how the neural signals combine to contribute to our own inner world, our qualia. But one must crawl before one walks, so even our limited knowledge of these neural systems will assist us towards the goal of emulating physiological systems with fidelity.

But other than the pleasure of gaining pure knowledge, why should we be interested in emulating biological systems? The answer is because we do what we do so very well. As a simple example, say one gives a stack of photographs to a typical adult, with the instructions to identify every photograph containing an image of any arbitrary automobile.

That person will effortlessly perform this task in minimal time, generally regardless of the characteristics of the automobile (size, pose, style) or the scene in which it is embedded. This is a remarkable trait, and one so natural we do not even think about. To instill such a capability into a machine would be of immeasurable benefit in areas as diverse as military automatic target recognition and medical physiological anomaly detection.

Obviously, more high-level neural processing is required for the above recognition task than a simple projection of dark and light patterns into the visual tract. For instance, using the car example, the subject must know what a car is, how it can vary and remain a car, what it looks like in different poses, and, in general, what defines that element of 'car-ness.' The visual perception process, then, must combine these elements with the representation of the object in the visual tract.

It thus becomes attractive to consider ways in which to construct a computational model capable of emulating this biological visual perception process, to include the ability to memorize the representation of an arbitrary object and to search for new instances of that object in an intelligent and efficient manner. The research reported on here explores the development and implementation of exactly such a capability.

1.2 Problem Statement and Scope

1.2.1 Problem Statement. This dissertation focuses on the development of a new computer vision model motivated by physiological and behavioral aspects of the human visual system. Such a model is shown to be capable of detecting, identifying, or verifying the presence of virtually any desired object within an arbitrary visual scene. Using this model, a search system will be developed that will be seen to memorize and locate arbitrary text or objects within a still image or a sequence of video images.

1.2.2 Scope. Though humans use numerous attributes of visual scenes in the process of object identification (e.g., spatial intensity gradients, presence of bars/edges, color, or motion), the characteristics used by the model developed here will be limited to non-temporal, intensity based attributes. That is, intensity gradients of oriented bars and edges within gray-scale, still images will serve as the primary features used for object identification. Note that though video imagery will be examined in the course of this research, it will be treated as a sequence of still images. Though the human visual tract contains cortical neurons responsive to color and motion, those traits will not be extensively explored here.

1.2.3 Research Contributions. In developing this vision model and search system, this research has made the following original contributions:

1. A new model of the human visual tract is introduced that emulates the function of neurons in the biological Retina, Lateral Geniculate Nucleus, and Primary Visual Cortex.
2. A view-based standardization process is developed that simulates the physiological characteristics of contrast sensitivity, contrast normalization, and cortical magnification to project an arbitrary visual field of view into a standardized, simulated cortical map.
3. A scanning algorithm is developed that in some sense emulates the saccadic behavior of the human visual system to permit complete exploration of arbitrary visual scenes.
4. The characteristics in 1-3 above are synergistically combined to produce a recognition model that is shown to be capable of recognizing arbitrary objects within arbitrary scenes.
5. A context-based recognition paradigm is introduced that solves the general threshholding problem encountered in pattern recognition processes.

1.3 Dissertation Organization

This dissertation is organized into seven chapters. Chapter II provides background on physiological and behavioral characteristics of the human visual system, and Chapter III provides motivation for using a behavioral vision model for object recognition. Chapter

IV introduces SCAN-IT, a new computer vision model based on these characteristics. Results from experimental trials are reported in Chapter V, and a demonstration of SCAN-IT's application to a real-world problem is provided in Chapter VI. Chapter VII summarizes the research and reviews the contributions made.

II. The Human Visual System

2.1 Introduction

For human beings, visual perception is a highly dynamic process, with multiple fixations required to explore scenes and identify (recognize) objects. These refixations, or saccades, are made necessary by the physiological make-up of the visual tract. The human visual system (HVS) actually has only a small region of high acuity centered about the point at which we are fixating. Saccadic behavior allows us to scan this small, high acuity region over relatively large spatial areas and thus collect and process multi-resolution data centered around multiple locations in the visual scene [42, 59]. The HVS has provided us the capability to ignore irrelevant data and collate the pertinent information obtained via the multiple saccades, producing a globally consistent, unambiguous view of the world. Part of this consistent world view is the illusion of uniform spatial resolution.

A brief introduction to the physiology and behavior of the human visual system would be informative at this point. A simplified description of the visual tract and how the visual field of view is formed will first be presented, followed by a discussion of saccadic behavior in humans. More detailed accounts of this physiology and behavior can be found in [17, 31, 32, 42, 43, 64, 77, 83]

2.2 The Physiology of Primate Vision

2.2.1 Early Visual Processing. Figure 1 shows a simplified view of the human visual tract. Light energy first enters the front of the eye and is focused by the cornea and lens on the retina at the back of the eye, centered around the *fovea* (see Figure 2a). Two fundamental types of photoreceptors exist in the human retina: rods and cones [77]. The rods are most responsive to low levels of light (scotopic), while the cones respond strongly to higher levels of light (photopic). The density of these receptors in the typical human retina is shown in Figure 2b, where it can be seen the density of cones is greatest in the foveal area, and rapidly decreases as the distance from the fovea increases. Rods, on the other hand, are not present in the fovea, but are distributed over a much larger area of the retina. It has also been found that the receptive fields of the photoreceptors tend to

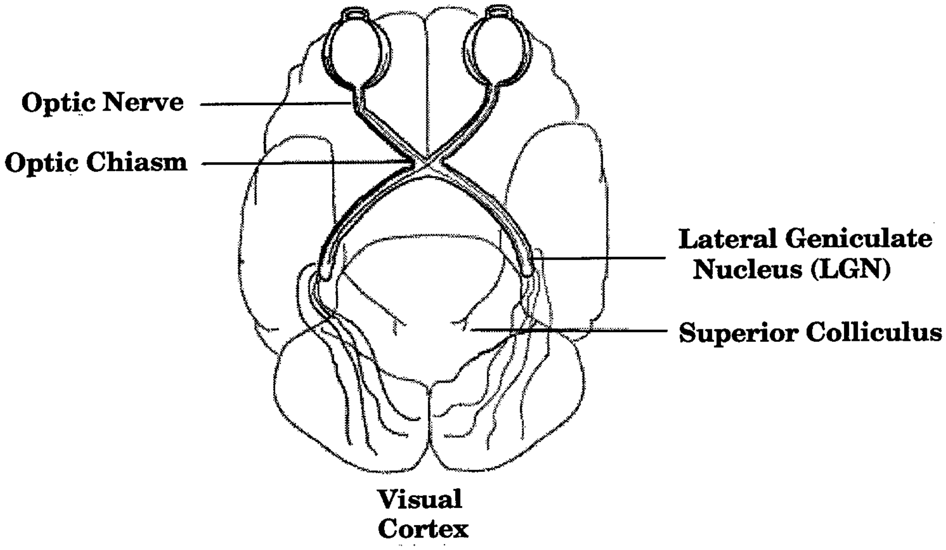
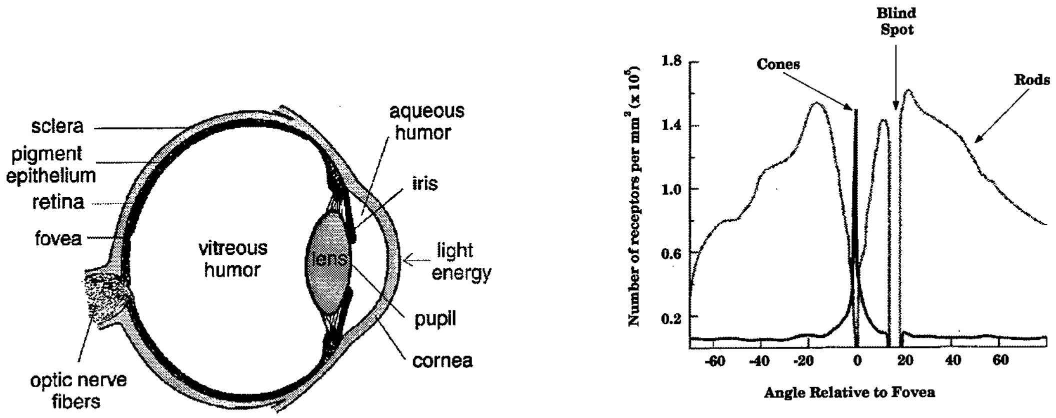


Figure 1. Simplified Diagram of the Human Visual Tract [70]

increase in size with increasing distance from the fovea [70, 77]. These facts suggest that under normal illumination conditions, the high density of cones in the foveal region will lead to a high resolution sampling of the portion of the visual field projected on that region; conversely, under low lighting levels, photoreceptor response in the foveal region will be quite poor, explaining why we are unable to distinguish dim light sources (such as weak starlight) while looking directly at them. The model constructed for this research assumes normal illumination, and thus a visual field predicated on resolution levels induced by the cone density.

Once the visual signals generated by the photoreceptors leave the eye, they pass through the *optic chiasm*, a location where the optic nerve outputs from the two retinae join and are then re-organized into two separate groups that encode information about the right and left visual fields. The majority of the retinal signals then enter the *Lateral*



a b

Figure 2. a) Human Eye. b) Density of photoreceptors in the retina. The cone receptors are concentrated in the fovea, accounting for greater resolution of parts of the visual scene projected on the fovea [77].

Geniculate Nucleus (LGN) layers, where the visual impulses are filtered and sent on to the *primary visual cortex* (V1) via the optic radiation pathway [1].

2.2.2 Primary Visual Cortex. Additional evidence for the multi-resolution field of view (FOV) can be found here in the visual cortex. Studies of the mapping of signals from the photoreceptors in the retina to neurons in the visual cortex have shown that more cortical area is devoted to processing the foveal signals than to the peripheral area signals, a phenomenon known as cortical magnification [6, 8, 44, 77, 78]. Moreover, the size of the neuronal receptive fields in the cortical map of the retina increases toward the periphery of the map, indicating reduced spatial sampling and a lower resolution peripheral cortical representation of the visual field. The bottom image in Figure 3 shows a slice through area V1 of a monkey, where the dark cells are those that were responding as the animal viewed the pattern in the upper image [70]. The three vertical lines in the tissue slice correspond to the half-rings on the right side of the stimulus, and one can see the response to each of the different-size rings encompasses virtually the same cortical area. The retinotopic

---

[1] The functions of the LGN are not currently well understood, but are thought to include enhancement of contrast information, organization of information (such as color or motion), and functioning as a site to receive and incorporate feedback from higher visual areas [70, 77]

organization on V1 is also clearly evident here; that is, specific portions of the visual scene projected on the retina are mapped to specific areas of the Primary Visual Cortex.

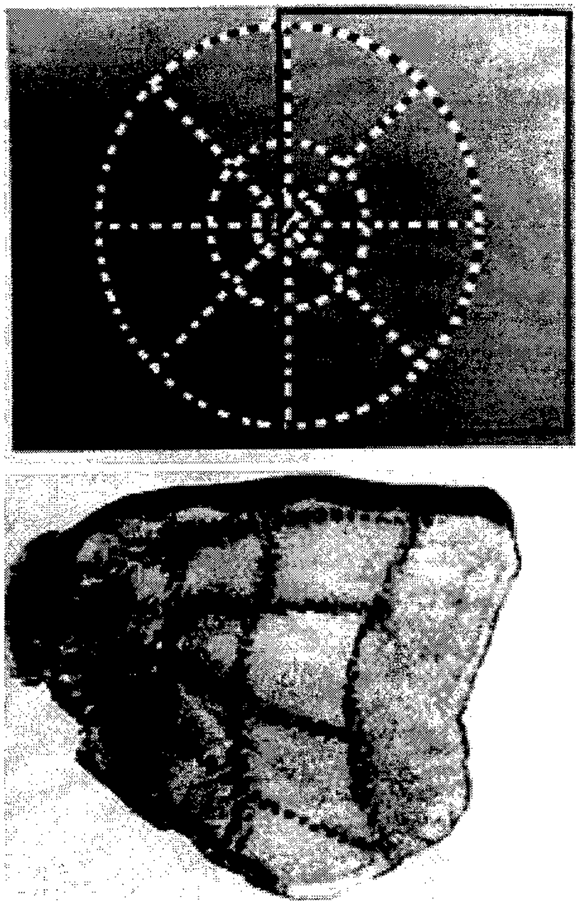


Figure 3. Cortical Magnification in V1. The bottom image shows a slice through area V1 of a monkey, where the dark cells are those that were responding as the animal viewed the pattern in the upper image [70].

Contrast sensitivity also plays a role in the construction of the visual field. Contrast is defined as

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} \tag{1}$$

where L_{max} is the maximum intensity of the stimulus and L_{min} is the minimum intensity [70]. Neurons in the visual tract can be characterized by the stimuli contrast needed to produce some criterion level of response to a harmonic function [72, 77]. When a contrast pattern is well suited to a neuron's receptive field, only a low level of contrast will be needed to produce the criterion response; an ill-suited pattern will require a greater contrast to produce the same response. The amount of contrast necessary to produce this response is called the contrast threshold, and its inverse is known as contrast sensitivity. Figure 4 shows an example of a typical human contrast sensitivity curve for photopic neurons with receptive fields covering foveal regions, where the bandpass peak occurs around eight cycles per degree [48, 77]. Though different test techniques and the selection of neurons in different portions of the retina may result in the curve shifting somewhat to the left or to the right, the bandpass nature remains similar to that seen here. For the vision model implemented in this research, the contrast sensitivity function shown in the figure will be used.

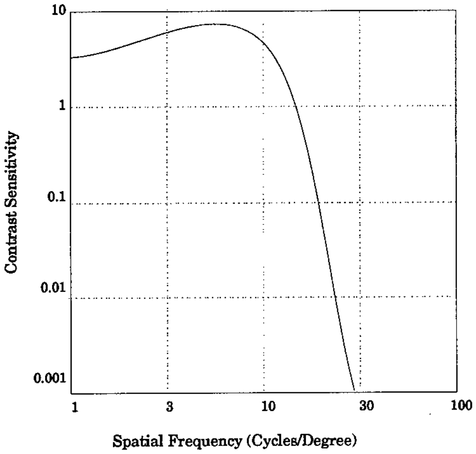


Figure 4. Typical contrast sensitivity function characterizing a photopic foveal neuron's response to harmonic stimuli [48, 77].

Orientation selectivity also appears for the first time in the visual cortex. The receptive fields of neurons in the prior portions of the visual tract (retina, LGN) have circularly symmetric receptive fields with center-surround properties (See Figure 5) [33, 77]. Figure 5 shows an **on-center, off-surround** receptive field which will cause neural excitation when light falls on the center, and inhibition when light falls on the surround. **Off-center, on-surround** neurons also exist, with the expected inhibitory/excitatory reactions. Because receptive fields of different sizes are present (with larger receptive fields tending toward portions of the visual field farther from the fovea), differently sized light (or dark) patterns

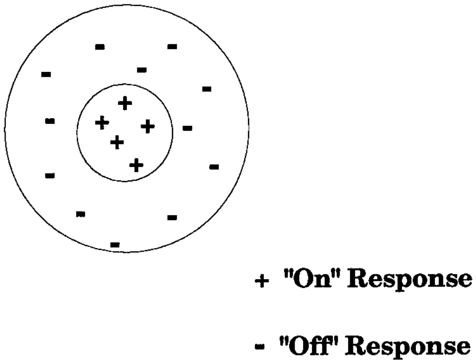


Figure 5. Center-Surround Organization of a neural receptive field in the visual tract [77].

will strongly excite or inhibit different neurons, but circular 'blobs' of the appropriate size will produce the most striking results.

In V1, for the first time neurons are found with receptive fields not necessarily circular or symmetric, but instead with a tendency to be oblong in nature with asymmetric on/off characteristics. Figure 6 shows three examples of such receptive fields, which will tend to produce excitatory responses to edges, dark bars (lines), and bright bars (lines), respectively. These receptive fields will be of multiple sizes and orientations, and will thus allow detection (neural activation) in the presence of bars and edges of different thicknesses (directly related to bar/edge frequency) and orientations. At this stage, then, the visual system is capable of responding to bars and edges at arbritrary orientations and multiple resolutions.

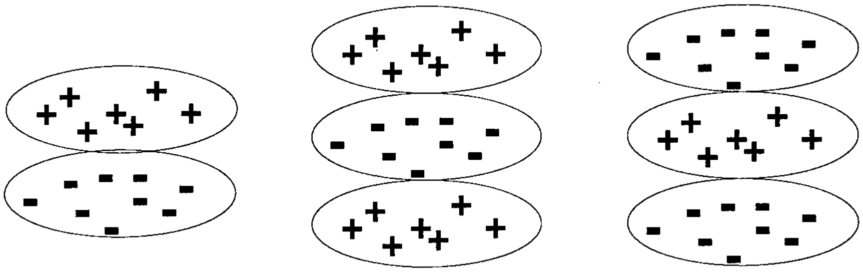


Figure 6. Non-circular neural receptive fields found in V1.

2.2.3 Other Spatial Visual Attributes. The above section describes the ability of neurons in the visual tract to process blobs and oriented bars/edges, but other spatially-based attributes also play a role in our ability to detect and interpret our visual field of view. The most obvious are probably motion and color, and brief descriptions of the neural bases for each are offered in the following two sections. However, we are able to easily perform recognition tasks with gray-scale, still imagery, suggesting that color and motion, though important, are not necessarily vital to all vision tasks. For that reason, this research has not delved particularly deeply into the use of color or motion for object recognition. The following sections are thus included for completeness, but the characteristics described are excluded from the model developed for this research.

2.2.3.1 Motion Sensitivity. In addition to detecting stationary bars/edges, neurons have also been found that specifically respond to spots, bars, or edges sweeping across the receptive fields, even to the extent of only being responsive to one direction of motion (right to left, for example) [31,32,77]. Figure 7 shows a model proposed by Sekuler to explain the directional sensitivity to motion [70].

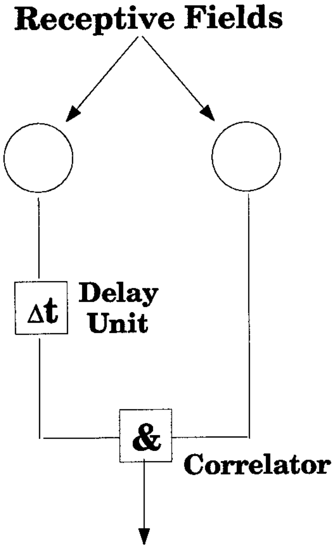


Figure 7. Model proposed by Sekuler to explain motion specificity of neurons [70].

As a stimulus passes through the receptive fields, responses are generated and passed down to the correlator, which will generate a neural output based on its input. The left path, however, has a time delay built in, so the neuron will maximally respond only when the stimulus first enters the left receptive field, and then after the interval Δt enters the right receptive field. The neuron will thus respond most strongly to a stimulus sweeping from left to right at the appropriate speed. Obviously, using this model one assumes that neurons exist with different Δt characteristics (and/or different spatial distances between the receptive fields) to allow response to stimuli sweeping across the fields at different speeds and directions.

2.2.3.2 Color Detection.

Color also plays a role in our ability to recognize visual scenes or identify specific objects within our field of view. The following discussion is generally based on descriptions given by Hall [24] and Martin [50], but more information on color physiology and processing can be found in [4, 33, 77, 83].

Our color detection capabilities again begin with the rods and cones lining the retina. These photoreceptors contain pigments that absorbs some wavelengths better than others, and researchers have found three distinct types of cones, differentiable by their spectral sensitivities to light (though the rods also contain pigment, this research assumes photopic illumination conditions, where only cones come into play). These three different types of cones act as bandpass filters, and are known as L (long wavelength peak), M (medium wavelength peak), and S (short wavelength peak) cones.

Signals produced by the cones are processed through neurons in the retina, and then are sent via the optic nerve to cells in the LGN. Three types of cells are found here: L cells, which are hyperpolarized by light stimulation regardless of the light's spectral composition; r-g cells which are hyperpolarized maximally by green and depolarized maximally by red; and y-b cells, which are hyperpolarized maximally by blue and depolarized maximally by yellow. These latter two cells are called color-opponent cells and appear to form two chrominance (color) channels to the visual cortex. The non-opponent L cells seem to carry brightness, or luminance, information about the field of view. The luminance and chrominance information is then carried to the visual cortex for processing into our inner perception of form and color.

2.3 Saccadic Behavior

The previous sections on motion and color sensitivity were included for completeness and described characteristics not to be incorporated into the research described in this document. Saccadic behavior, on the other hand, will play an integral role in the model developed here. An example of this behavior is shown in Figure 8. Figure 8a shows a photograph of a girl, while 8b shows a typical recording of the eye movements made by a human test subject while exploring the image [81]. The subject was instructed to examine the picture with both eyes, and the plot shows the scan path recorded over a one minute period. One notes that a preponderance of the fixations are made to the eyes, nose, and mouth of the girl, hinting that these areas somehow provide added attraction to the attentional focus. Indeed, several researchers have shown that the calculation of saccadic targets is highly dependent on the particular task at hand (memorization or search, for example) and the familiarity of the object/scene being examined [2,29,43,81]. A saccade

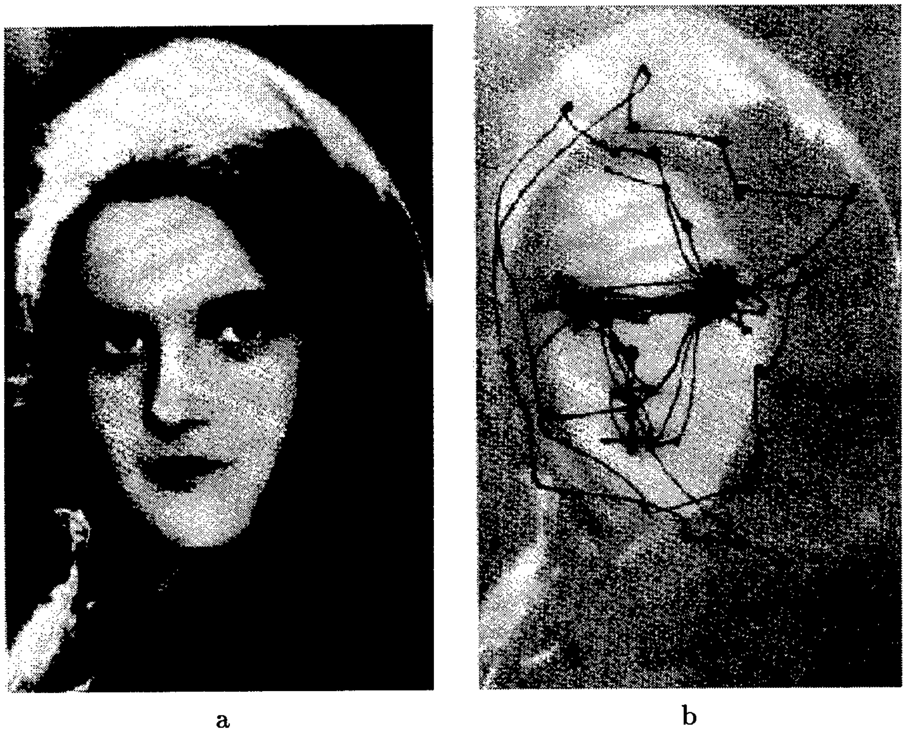


Figure 8. a) Photograph of a girl b) Recording of eye movements during examination of the photograph [81]

is thus made to the region containing the most interesting features, the features that will best solve the current problem.

During visual exploration or recognition tasks, the human eyes will fixate on the portions of the visual field which have been deduced to have the most information (relative to the particular task being performed), but simultaneously the lower resolution portions of the visual field will be explored to determine the next relevant location on which to fixate. These two different functions are known as *attention* (fixating on or attending to a particular image region) and *pre-attention* (determining the next attentional region). Neural implementations of these processes, and how they cooperate to give us our unambiguous view of the world, are not very well understood at this time, despite the large amounts of ongoing research (see, for example, [16, 21, 34, 40, 59, 68]).

Though not universally accepted, several researchers have supported the importance of the scanpath (sequence of fixations) in visual perception and recognition [14,59,61,62,81]. Noton and Stark compared the scanpaths of human eye movements in image memorization and subsequent image recognition, and found the paths to be topologically similar [55][2]. They suggest that an object is memorized and stored as an alternating sequence of object features and eye movements required to attend to the next fixation point; that is, that motor memory seems to be a vital part of recognition. The recognition process, then, consists of the appropriate eye movements followed by verification of the expected image features at each fixation point.

Other neuro-anatomical and psychological data have supported this behavioral pattern [41, 53, 75]. It has been proposed that two major pathways seem to be present in the higher levels of the human visual system, dubbed the 'where' path, involved in representing spatial and spatial relationship information, and the 'what' path, dealing with representing object features. These paths appear to provide the link between the low-level, retinal processing and the high-level brain structures involved in visual perception.

---

[2] This is not to imply that the scanpath (ordered sequence of saccades) accomplished during examination of an object or scene will be identical over multiple scans, but the paths can be expected to be similar.

2.4 Summary

In this section background has been provided on certain physiological and behavioral aspects of the primate visual tract. Though massive amounts of research have been accomplished (and continue to be accomplished) on virtually every aspect of visual processing, the concentration here has been on providing an overview of the characteristics with which the vision model developed for this research will be imbued. The next chapter will provide motivation for using a behavioral model for object recognition and will provide implementation details of the existing computational foundation upon which the new vision model is based.

III. Why a Vision Model Motivated by Behavior?

3.1 Introduction

The model developed during this research relies on behavioral aspects of the human visual system, but of what benefit is using such a model for object recognition in natural or complex scenes? Perhaps the most persuasive argument is that our own visual perception system performs remarkably well using such behavior. Though admittedly we are still far from a complete and precise description of visual processing, having a proven, working biological system naturally motivates the emulation of that system in a computational (in the sense of non-biological computation) setting.

Rimey and Brown state in [63] that "a task-oriented system should perform its task with the minimum necessary effort." The emphasis on extracting features from only a small portion of the visual scene at a time meets this objective. Extraction of global features from a scene containing the object to be recognized within a natural background will be computationally expensive and memory intensive. The scene background will supply many distractors to the object, and the recognizer must be capable of discounting this interference between features of the object and features of the background (the so-called *Feature Integration* and *Binding* problems). Also, the recognizer must be invariant to spatial variations (scale, translation, and rotation), in most cases requiring potentially extensive global transformations to an alternate domain.

Using saccadic behavior, storage requirements for the representation of each field of view (FOV) can be small, allowing quick comparisons (error measure calculations) of memorized FOVs with those extracted from new scenes. Because the FOV is small when compared to the entire image, the background is likely to play only a minor role when the focus of attention mechanism locates a familiar region on the memorized object. Once this familiar FOV is found, the memorized scanpath can be repeated over the new scene (in a spatially invariant manner, as explained in Chapter IV), and at each fixation the comparison will still be a computationally simple one. Evidence will then be accumulated at each fixation point that will either support or detract from the overall hypothesis that the object has been found.

Though much behavioral research has been accomplished in the field of active vision (e.g., [9, 17, 19, 21, 27, 43, 60]), few models have been proposed to serve as an engine for object recognition systems. Rimey and Brown use Hidden Markov Models [62] and Bayes Nets [63] to determine the fixation sequences that will best provide evidence that a specified activity or task is present within a visual scene (for example, a table place setting identified as *fancy* or *non-fancy* but does not *per se* recognize single objects via multiple views. Hecht-Nielsen presents a neural net active vision system that uses a saccadic process to find possible targets within a scene [28]. Probably most similar to the model resulting from the research reported here is one described by Rybak in [66, 67], where an artificial visual field is constructed and edge features extracted at 48 specified spatial locations. Saccadic behavior is then incorporated based on the competing attraction of those points. Because SCAN-IT has this model at its roots, it is worthwhile to provide additional details on the Rybak implementation.

3.2 Rybak's Saccadic Vision Model

3.2.1 Top Level Overview. The first stage of Rybak's model involves memorizing a segmented object via a sequence of fixations and saccades. Edge features at each fixation are memorized and stored in Sensory Memory, and the scanpath of the saccades is stored in Motor Memory. Once the system has learned what the object looks like, another image containing the object is presented. A search algorithm is used to scan the image, and when a fixation results in a match with a memorized feature set, a hypothesis is formed that the object has been found. At this point, the motor memory is accessed, and the scan path used during object memorization is invoked. If the expected feature sets are found at the expected locations, the hypothesis is accepted. Otherwise, the hypothesis is rejected and the system reverts to search mode.

Figure 9 shows a block diagram of the Rybak model. The Attention Window first transforms the image into a multi-resolution retinal image, followed by primary feature detection on the retinal image. The feature detector consists of orientationally selective neurons to detect edges and their orientations. The basic edge is defined as the edge located at the fixation point, and all other edges are defined as context edges. Rotational

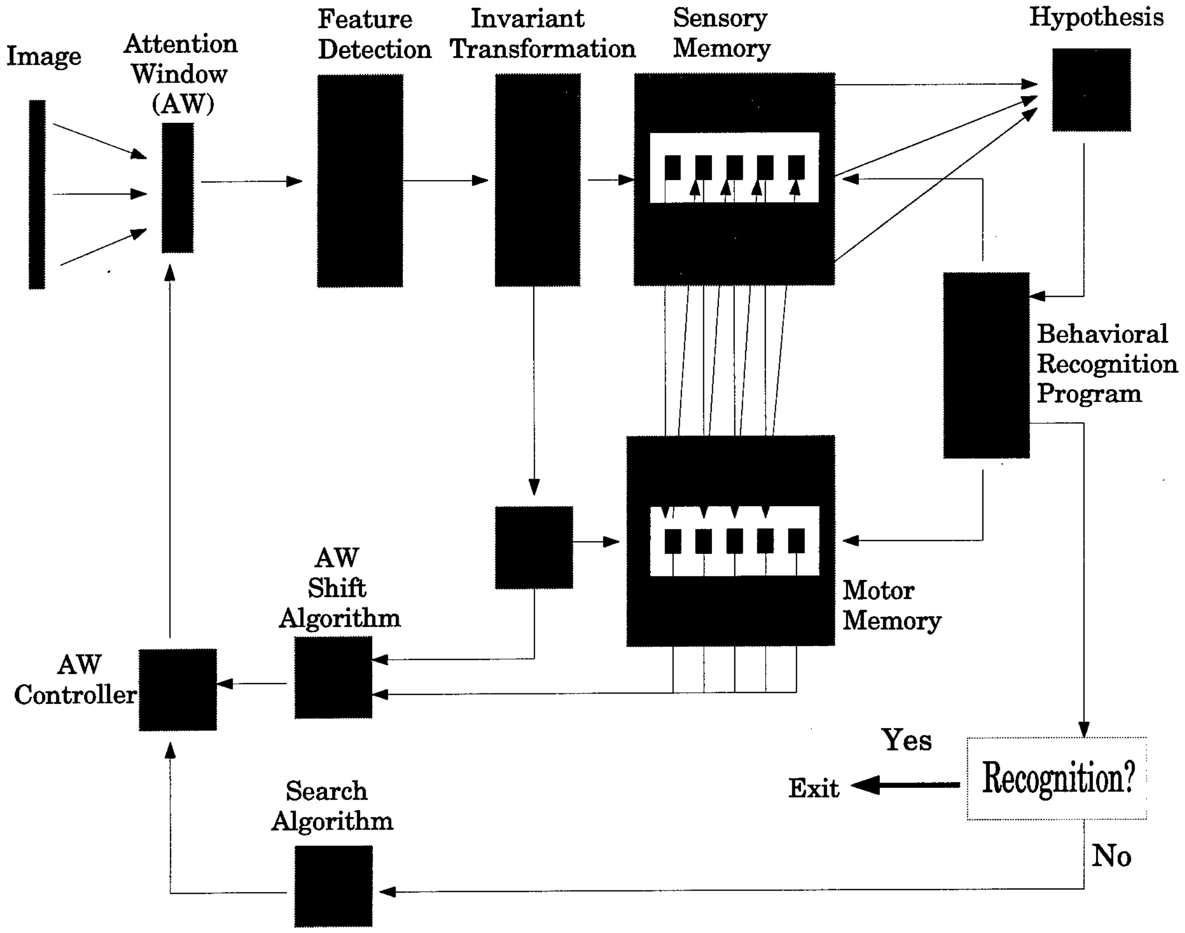


Figure 9. Block Diagram of the Rybak Visual Perception Model [66]

invariance is then provided by a coordinate transformation based on the intensity gradient of the basic edge. One of three different modes is then selected:

Memorization. Image patch features are extracted at a sequence of fixation points and stored in sensory memory. The next fixation point is determined by the attraction of each of the context edges. The successive saccades are stored in motor memory, and the object is then defined by the alternating sensory and motor memory traces.

Search. Under control of a search algorithm, an image is continually scanned until a retinal image similar to one of the stored images is found (in sensory memory). At this point, the hypothesis that the object has been found is formed and the system invokes the recognition mode.

Recognition. Using the motor memory data obtained from the model object, the system attempts to verify the data recalled from the sensory memory for each memorized fixation point. If a sequence of successful matches occurs, the object is recognized. Otherwise, the system returns to search mode

3.2.2 Attention Window (AW).

The major function of the AW is to transform the initial image patch into a retinal image with resolution that decreases from the center to the periphery. Rybak accomplishes this using a Gaussian-like convolution proposed by Burt [5]. Given the initial image $I = \{x_{ij}\}$ and resolution level l, then the value of each point in the image patch is calculated by

$$x_{ij}^{l} = \sum_{p=-2}^{2} \sum_{q=-2}^{2} g_{pq} \cdot x_{i-(2^{l-1} \cdot p), j-(2^{l-1} \cdot q)} \ , l = \{0, 1, \cdots, L\}$$

where L is the lowest resolution level to be calculated and g_{pq} is defined as

$$[g_{pq}] = \frac{1}{256} \cdot \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} , \ p \text{ and } q = \text{-2, -1, 0, 1, 2.}$$

The transformation from the initial image $I = \{x_{ij}\}$ into the retinal image $I^{R}(n) = \{x_{ij}^{R}(n)\}$ then must be calculated for each nth fixation point. Given three levels of resolution centered at fixation point $(i_0(n), j_0(n))$ where resolution is l_0 (the resolution of the original image), the fixation point is the center point of three concentric circles with radii defined by

$$R_0(l_0) = 1.5 \cdot 2^{l_0} \ ,$$
$$R_1(l_0) = 1.5 \cdot 2^{l_0+1} \ ,$$
$$R_2(l_0) = 1.5 \cdot 2^{l_0+2} \ .$$

The retinal image at the nth fixation point is then determined by

$$x_{ij}^{R}(n) = \begin{cases} x_{ij}^{l_0(n)} & ,\text{if } \rho_{ij}(n) \leq R_0(l_0) \\ x_{ij}^{l_0(n)+1} & ,\text{if } R_0(l_0) \leq \rho_{ij}(n) \leq R_1(l_0) \\ x_{ij}^{l_0(n)+2} & ,\text{if } R_1(l_0) \leq \rho_{ij}(n) \leq R_2(l_0) \end{cases}$$

where

$$\rho_{ij}(n) = \sqrt{(i - i_0(n))^2 + (j - j_0(n))^2}$$

The retinal image thus has the highest resolution at the center region, next highest within the annulus surrounding the central circle, and lowest within the outer annulus. Figure 10 shows the areas of different resolution within the AW.

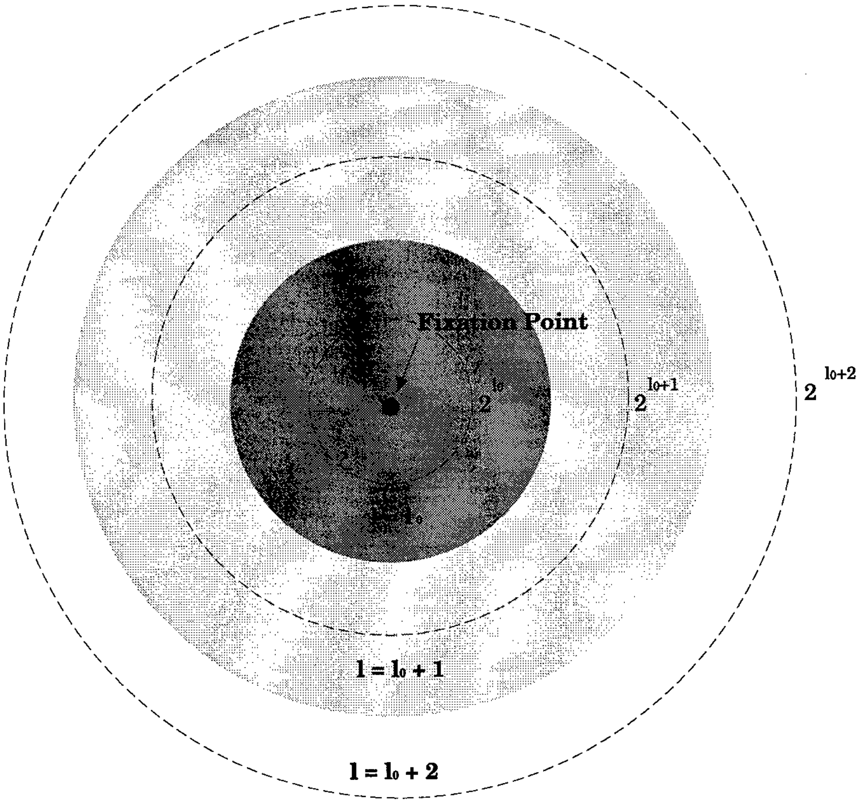


Figure 10. Multiple Resolution Annuli Surrounding Foveation Point

3.2.3 Feature Detection. Hubel and Wiesel pioneered the theory that the primary visual cortex contains frequency and orientation selective neurons [30–32], which in the simplest case detect the arbitrary orientation of edges at each point in the retinal image. In the Rybak model, such edges are considered to be the primary features, and are detected with resolution dependent on their location within the visual field. The orientationally dependent receptive field (ORF) is implemented using 16 oriented neurons based on the description by Grossberg, *et al* [21], as the difference between two Gaussian convolutions with spatially shifted centers. Given the points (p, q) in the visual field (where $p, q \in \{-2, -1, 0, 1, 2\} \equiv S$), the input to a neuron tuned to orientation α at coordinate (i, j) is given by

$$Y_{ij\alpha} = \sum_{pq \in S} x^R_{pq} \cdot (G^+_{pqij\alpha} - G^-_{pqij\alpha}) \ ,$$

where

$$G^+_{pqij\alpha} = e^{-\gamma^2 \cdot ((p-i-m_\alpha)^2 + (q-j-n_\alpha)^2)},$$
$$G^-_{pqij\alpha} = e^{-\gamma^2 \cdot ((p-i+m_\alpha)^2 + (q-j+n_\alpha)^2)}$$

with γ the reciprocal of the variance. m_α and n_α are determined from the ORF orientation as

$$m_\alpha = d(l) \cdot \cos(2 \cdot \pi \cdot \alpha/16),$$
$$n_\alpha = d(l) \cdot \sin(2 \cdot \pi \cdot \alpha/16).$$

where $d(l)$ is the distance from the center of each Gaussian to the center of the ORF, and is found by

$$d(l) = \max\{2^{l-2}, 1\}$$

The 16 co-located neurons, each with a different orientation, undergo a competitive (inhibitory) interaction to achieve steady-state equilibrium, obeying the system of equations

$$\tau \cdot \frac{dV_{ij\alpha}}{dt} = -V_{ij\alpha} + Y_{ij\alpha} - T - b \cdot \sum_{k=0}^{15} Z_{ijk}, \ k \neq \alpha$$

$$Z_{ij\alpha} = f(V_{ij\alpha}); \ \ \alpha = 0, 1, 2, \cdots, 15 \ ,$$

where $V_{ij\alpha}$ and $X_{ij\alpha}$ represent, respectively, the neuronal membrane potential and the output of the neuron at location (i, j) with orientation α. The parameter b characterizes the reciprocal inhibition, T is the neuronal threshold, and τ is the time constant associated with the neuronal threshhold. The nonlinear function $f(V)$ is defined by

$$f(V) = \begin{cases} V & ,V \geq 0 \\ 0 & ,\text{otherwise} \end{cases}$$

The system of differential equations achieves equilibrium at one of two solutions:

- all $Z_{ij\alpha} = 0$ (if all $Y_{ij\alpha} < T$)
- only one $Z_{ij\alpha} = Y_{ij\alpha} - T > 0$ at orientation ϕ, and $Z_{ij\alpha} = 0$ if $\alpha \neq \phi$.

Then an edge is considered to exist at (i, j) for the second case, with orientation ϕ and brightness gradient $Z_{ij\phi}$; an edge is not detected for the first case.

At each fixation point, edge detection takes place at 49 locations on the retinal image: at the fixation point itself (the basic edge), and at the intersections of 16 radiating lines (corresponding to the orientation of the 16 neurons) with the three concentric circles defining the regions of different resolutions (see Figure 10). Thus, 48 context edges may be defined for every fixation point.

3.2.4 Invariant Transformation. The coordinate system is next transformed into a feature-based frame of reference. A new reference axis is defined along the vector of the brightness gradient of the basic edge, and thus each context edge can be invariantly represented in this relative coordinate system by three parameters ϕ, ψ, and λ (see Figure 11). The parameter ϕ is the relative orientation of the context edge (with respect to the basic edge), ψ is its relative angular location, and λ defines the distance from the fixation point. These parameters are calculated as follows:

$$\phi = \text{mod}_{16}(\phi_c - \phi_0 + 16);$$
$$\psi = \text{mod}_{16}(\psi_c - \psi_0 + 16);$$
$$\lambda = l - l_0$$

Thus each retinal image can be invariantly represented by three, 16-component vectors:

$$\vec{\phi}_{\psi\lambda}(n) = \{(\phi_0(n), \phi_1(n), \cdots, \phi_{15}(n))_\lambda : \ \lambda = 1, 2, 3\}$$

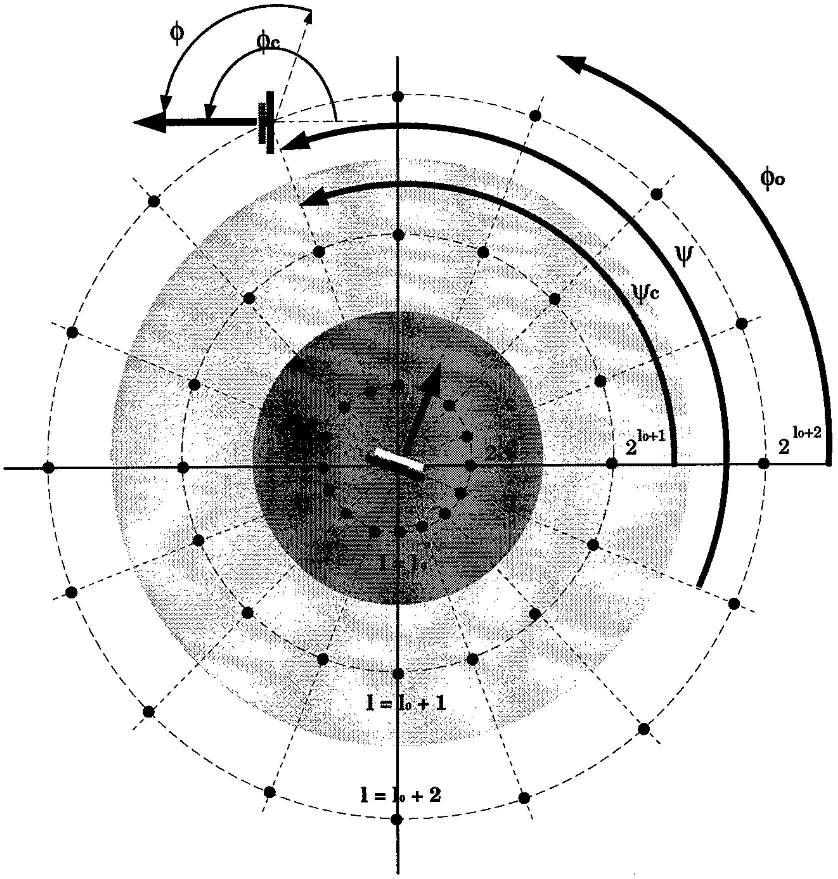


Figure 11. Relative Coordinate System with Respect to Basic Edge [66].

These vectors are stored in the sensory memory for comparison with those calculated from new retinal images. One similarity metric proposed by Rybak is

$$D(\vec{\phi}_{\psi\lambda}, \vec{\phi}^*_{\psi\lambda}) = \frac{1}{N}\sum_{\lambda}\sum_{\psi}\frac{1}{1+8\sin^2\left(\frac{\pi}{16}\cdot(\phi_{\psi\lambda}-\phi^*_{\psi\lambda})\right)}$$

where N is the number of edges in the retinal image. The two images $\vec{\phi}_{\psi\lambda}$ and $\vec{\phi}^*_{\psi\lambda}$ are considered similar if D exceeds an experimentally determined (and application or database dependent) threshold.

3.2.5 Selecting the Next Fixation Point. Each context edge in the current retinal image is considered a potential target for the next fixation point. The shift from the nth fixation point to the n+1th point is stored in the motor memory as a set of shift parameters, calculated relative to the nth coordinate system. These parameters are defined as

$$\begin{aligned}
\psi_{\text{shift}}(n, n+1) &= \psi_{n+1}(n) \\
l_{\text{shift}}(n, n+1) &= \lambda_{n+1}(n) \\
\theta_{\text{shift}}(n, n+1) &= l_0(n+1) - l_0(n)
\end{aligned}$$

where ψ_{shift} and l_{shift} define the relative direction and distance of the shift, and θ_{shift} defines the change of resolution in the central region of the AW when attention moves from the current fixation point to the next (for the current implementation θ_{shift} will be 0).

The selection of the next fixation point is a non-trivial problem, and is generally accepted to be task-dependent, relying in humans on some abstract cognitive processing. In the Rybak model, each context point attracts the AW with a force A_k, defined as

$$A_k = A_1 \cdot \frac{Z_k}{Z_{max}} + a_2 \cdot \frac{\lambda_k}{2} + a_3 \cdot \eta_{ij}(n) + a_4 \cdot \chi_{ij}.$$

The first term on the right hand side of the equation provides for the dependence of the attraction on a normalized value of the brightness gradient in the context point; the second term provides for the dependence of the attraction on the relative distance of the context edge from the current fixation point; the third term is included to prevent cycling repeatedly back to the same area. The function $\eta_{ij}(n)$ determines the novelty of the area containing the selected context point, and is set to zero immediately after selection for all points within some distance of the new fixation point. The function is then allowed to recover to a value of 1 over time. The parameter χ_{ij} defines a semantic significance of the area including the context point, and may be defined appropriately if knowledge is available (for example, if the object being memorized is a human face, additional attraction can be invoked by context points in the vicinity of the eyes, nose, and mouth [81]. The coefficients a_x are selected experimentally during training.

3.2.6 Results and Limitations. Rybak showed that the system was able to memorize, then recognize objects in gray-level images invariantly with respect to shift, rotation, and scale. However, certain limitations are imposed by the model that decrease both its effectiveness and its fidelity (with human biology).

First, the cortical receptive fields generated by Burt's Gaussian-like convolution do not have a very accurate relationship with those found in the primate visual cortex [10,12, 79]. Other functions exist that can be used to better represent the oriented receptive fields of simple cells, and in the next chapter we propose the Gabor as an appropriate option.

Next, it is agreed that the saccadic behavioral process as described in Section 2.3 is an appropriate method to emulate Mishkin's 'where' path, but what are the desired features to be processed along the 'what' path? As discussed in Chapter II, it is well accepted that the visual pathway contains cells uniquely responsive to different stimulus dimensions, such as frequency, bar/edge orientation, bar/edge size, bar/edge direction of motion, color, and others. It would seem logical then, that the set of such features, or a subset thereof, would be appropriate for the implementation of a physiologically-based visual system. Olshausen [57] and Bell and Sejnowski [3] propose that edges and bars are actually the independent components forming a basis set for natural images, and that it is thus appropriate and natural that the human visual system contains edge and bar detectors. Thus, though biological neurons exist that are responsive to the other stimulus dimensions, it appears that those responsive to edges and bars should be adequate to perform memorization and recognition in natural scenes. The Rybak model indeed uses edges as features, but a more effective model will be one based on detection of both bars and edges.

A major limitation of the Rybak model is the arbitrary restriction to 48 assigned (by location) points for both feature extraction and saccadic targeting. There is very little guarantee that the interesting features will always be located at those points surrounding the fixation, and indeed, much more information is lost than retained using such a method. Tighter segmentation is also required prior to memorization, as any features extracted from non-object background will simply add noise to the final representation of the memorized object. During saccadic memorization, because a saccade can only be made to one of those 48 locations, it will tend to take a large number of saccades to rediscover the object in a new setting. Indeed, in a local implementation of the model, though recognition of simple objects was adequately performed, greater than 200 saccades were required to locate those objects in images sized to 128 × 128 pixels.

3.3 Summary

This chapter has motivated the use of behavioral and physiological characteristics for computational vision, and has provided a detailed description of the vision model proposed by Rybak, on which the model developed during this research is based. The limitations of the Rybak model (poor receptive field representation, inadequate feature selection, and restrictive spatial limitations) have been discussed, and the ways in which they will be overcome are addressed in the next chapter, where a new computational vision model is introduced.

IV. SCAN-IT: A New Model of Computational Vision

4.1 Introduction

This chapter introduces SCAN-IT, Saccadic Control for Active Neural Identification of Targets. SCAN-IT is based on physiological and behavioral characteristics described in Chapter II and is designed to emulate, in some sense, the active vision process as exhibited by humans.

The chapter will begin with a top-level overview of SCAN-IT. Implementation details are then discussed, beginning with a description of the Gabor function and its suitability in emulating a neural receptive field. Construction of a simulated visual field of view (FOV) is described, as well as the transformation of this FOV into a view-standardized space and finally to a representative feature matrix. Function of the model in memorization mode is presented in detail, as is the function in search and recognition modes.

4.2 SCAN-IT Overview

The vision model implemented here operates in three fundamental modes: memorization, search, and recognition. During memorization, a fixation point is automatically selected on the object/scene being memorized, the visual FOV around that point is constructed, and the features within the FOV are memorized (saved as a feature matrix). A new fixation point (saccadic target) is then automatically selected, the focus of attention moves to the new location, and the entire process repeats for the user-specified number of saccades. This scanning process results in an alternating sequence of feature matrices and saccadic instructions (describing the direction and distance from the previous fixation point) that represents the object/scene of interest.

In the search task, a new image is presented to the model, and a saccadic scan is performed. At each fixation, the visual field is again constructed and the features extracted, but rather than simply saving the data, this feature matrix is compared with each of the feature matrices computed during memorization. If a match is found (where the meaning of "match" will be discussed in Section 4.6), the hypothesis is formed that indeed the object/scene is present in the new image. At this point, recognition mode is invoked, and the sequence of saccades recorded during memorization is repeated over the new image. If the expected features are found around each fixation point, the object/scene is recognized, otherwise the model reverts to search mode. If the image has been fully explored with no successful recognition, the search is terminated with the conclusion that the object/scene is not present in the new image.

4.3 SCAN-IT's Visual Field Representation

4.3.1 The Gabor Filter for Oriented Bar/Edge Detection. In 1980, Daugman proposed a novel model for the two-dimensional receptive field of simple cells in the visual cortex [10]. This model was based on a family of functions first introduced by Gabor [18] which were composed of one-dimensional sinusoids constrained by one-dimensional Gaussians. Daugman's work introduced the concept of two-dimensional Gabor functions, comprised of sinusoidal planar waves constrained by Gaussian envelopes[1]. Though all spatial-frequency filters have an inherent space/frequency trade-off, filters built using Gabor kernels have been shown to provide optimal joint resolution in the spatial and frequency domains [10, 13, 18]. The spatial domain impulse response function $G(x, y)$ of a two-dimensional Gabor filter is given by [11]

$$\begin{aligned} g(x, y) &= \exp(-\pi[(x - x_0)^2\alpha^2 + (y - y_0)^2\beta^2]) \\ &\quad \cdot \exp(-2\pi j[u_0(x - x_0) + v_0(y - y_0)]) \end{aligned} \tag{2}$$

which is characterized by the position parameters (x_0, y_0), modulation (frequency) parameters (u_0, v_0), and scale (spread) parameters (α, β). The function $g(x, y)$, then, is the product of an elliptical Gaussian centered at (x_0, y_0) with aspect ratio β/α, multiplied by a complex modulated exponential with spatial frequency $\sqrt{u_0^2 + v_0^2}$ and orientation $\tan^{-1}(v_0/u_0)$. The Fourier Transform of $g(x, y)$ is given by

$$\begin{aligned} G(u, v) &= \exp\left(-\pi\frac{(u - u_0)^2}{\alpha^2} + \frac{(v - v_0)^2}{\beta^2}\right) \\ &\quad \cdot \exp\left(-2\pi j[x_0(u - u_0) + y_0(v - v_0)]\right) \end{aligned} \tag{3}$$

---

[1] The mathematical foundations of Gabor theory and its applicability to cortical modeling are presented in detail elsewhere [10, 11, 13, 18, 20, 37, 74].

The analytic functions described above contain both even-symmetric (cosine) and odd-symmetric (sine) parts in complex space. To describe a physically meaningful quantity that can be related to the receptive field response profile of a biological neuron in the visual cortex, one would like to create a real filter profile with the desired symmetry, modulation, and orientation properties. A filter to detect the illumination gradients of oriented edges at a specified frequency can thus be constructed using the pure sine component of $g(x,y)$ (see Figure 12a), and the associated Fourier Transform $\frac{1}{2}\{-jG(u,v)+jG(-u,-v)\}$ (see Figure 12b). Oriented bars can be similarly detected using the pure cosine component of $g(x,y)$, $\frac{1}{2}\{G(u,v)+G(-u,-v)\}$ (Figure 12c and d). For the functions shown in Figure 12, $x_0 = y_0 = 0$, $v_0 = 0$, $u_0 = \pm 8$ cycles/degree, and $\alpha = \beta = \sqrt{(\pi/16)}$ deg^{-1}.

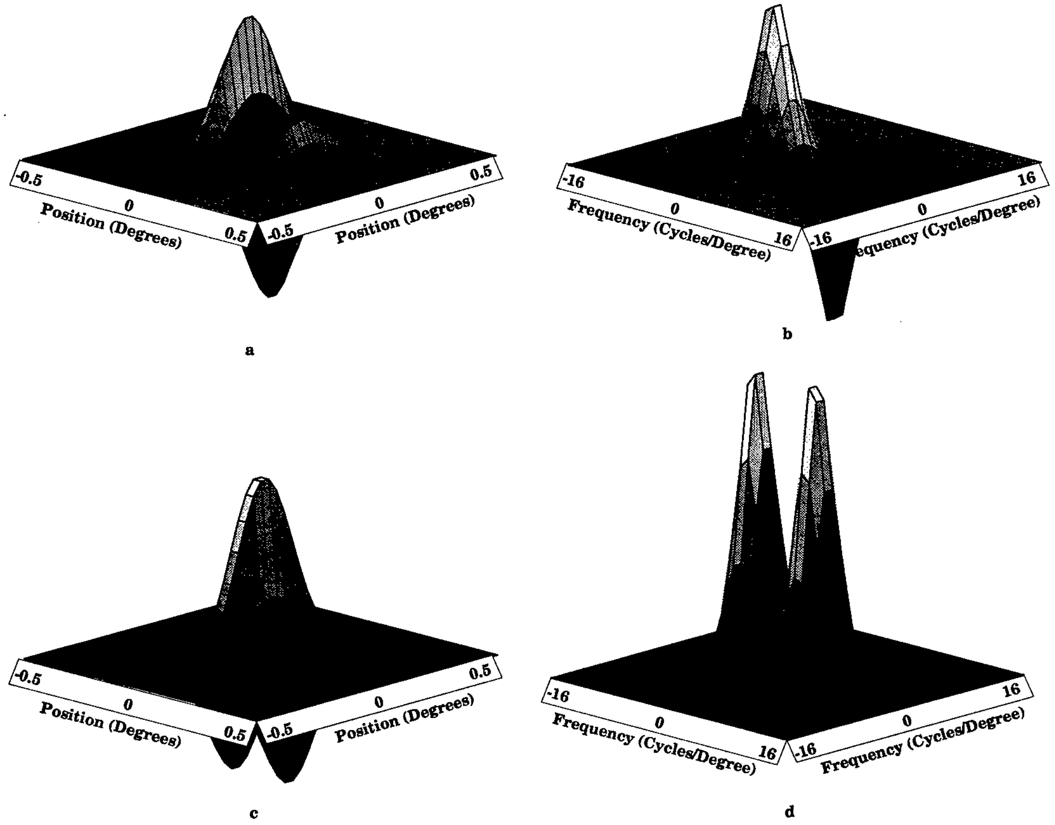


Figure 12. Two dimensional Sine and Cosine Gabor functions. (a,c) Spatial domain response (b,d) Fourier transform of Gabor function.

In 1984, Jones and Palmer performed direct measurements of cortical cells in a cat's visual cortex and demonstrated that the Gabor model indeed approximates Simple Cell receptive fields [37]. Zeki theorizes the cells in the primary visual cortex are organized to form multiple views of the retinal image, where the cells are selective to particular spatial frequencies and orientations [82]. The multiple views can be implemented by processing an image through multiple Gabor filters, each tuned to a unique spatial frequency and orientation. As shown in Figure 12b, a single 2D Gabor function can be constructed to cover two anti-symmetric, elliptically shaped regions in the frequency domain, and through the use of multiple filters, a broad range of spatial frequencies and orientations can be detected. In other words, an ensemble of such filters will be capable of detecting edge gradients at different resolutions and different orientations within an artificial visual field. Similarly, detection of oriented bars can be accomplished using an ensemble of filters based on the response seen in Figure 12d.

4.3.2 Constructing the Visual Field of View (FOV). The FOV in this model is represented by four separate mappings, two representing edge/bar magnitudes, and two representing edge/bar orientations at the associated spatial locations. The circular regions resulting from these mappings decrease in resolution as distance from the center, or fixation point, increases. Figure 13 gives a block diagram of the FOV building process, and Figure 14 provides a more detailed example of the construction of a multi-resolution edge intensity map. The orientation maps will be the same shape and size, but the entries will be the orientation of the dominant edges/bars at that location, as explained below.

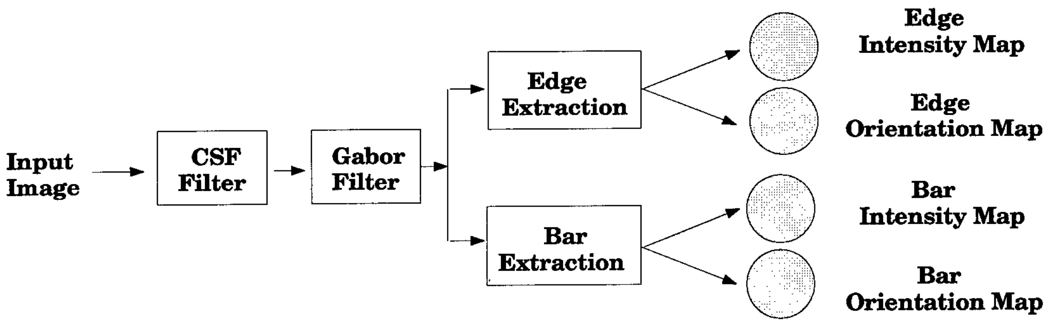


Figure 13. Block diagram showing the construction of the Field of View.

The image under examination is first bandpass filtered (using a functional form proposed by Mannos [48]) to emulate the responses of biological neurons to contrast variations

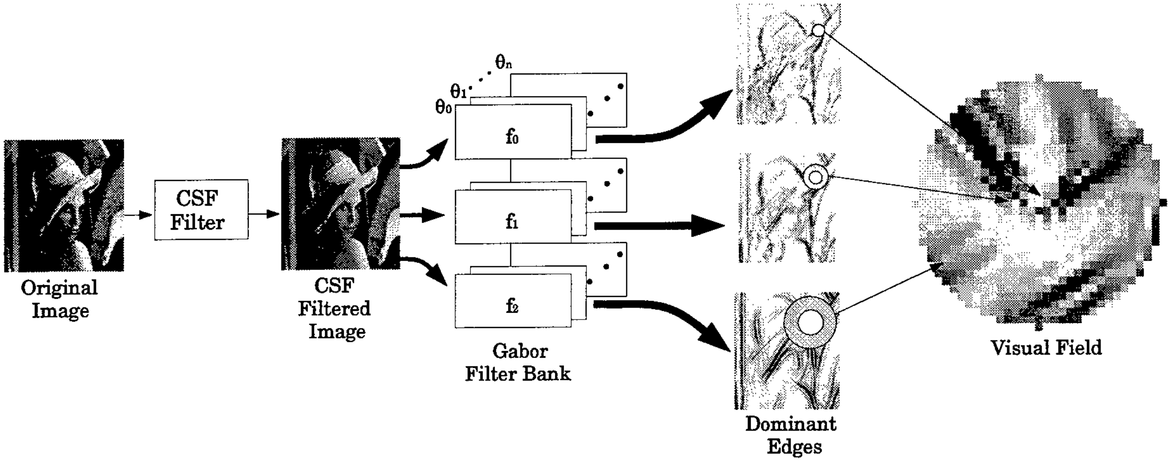


Figure 14. Example of the construction of the edge intensity map.

in the visual scene. The output is then processed through a filterbank consisting of multiple Gabor filters parameterized at different frequencies (resolutions) f and orientations θ:

$$\mathcal{R}(x,y,f,\theta) = I(x,y) * g(x,y,f,\phi), \quad f = \tfrac{N}{2^l},\ l = 0 \cdots L-1;\ \phi \in \Phi \qquad (4)$$

where $*$ indicates convolution, $\mathcal{R}$ are the filter outputs, I is the original image, g is the spatial domain Gabor function, N is the Nyquist frequency of the original image being processed, L is the maximum number of resolution levels, and Φ is the set of all desired orientations. As in Equation (2), f and ϕ are parameterized by v_0 and u_0 as $f = \sqrt{(u_0^2 + v_0^2)}$ and $\phi = \tan^{-1}(v_0/u_0)$.

For the work presented here, three resolution levels are used ($L = 3$), with 12 orientations, in 30-degree increments, at each level ($\Phi = \{30, 60, \cdots, 360\}$). Such an orientation distribution correlates well with that found in physiology [32], and results in a total of 72 filtered images from the filterbank, representing the responses to edges and bars at 36 different orientation/resolution combinations. First for the edge images, and then the bar images, a Max operation is invoked at each resolution level to determine, at every spatial location (every pixel location), the orientation producing the response with the greatest magnitude:

$$\mathcal{R}_{max}(x, y, f) = \max\{\mathcal{R}(x, y, f_i, \phi)\ ,\ i = 0 \cdots L-1;\ \phi \in \Phi\} \tag{5}$$

This process will produce a magnitude edge map, a magnitude bar map, and the two corresponding orientation maps at each resolution level (only the edge magnitude map is shown in Figure 14), representing the maximally responding orientations at every point in the image. The artificial visual fields $\mathcal{F}_{Mag}(x, y)$ and $\mathcal{F}_{Orient}(x, y)$ can then be constructed by assembling the appropriate disk/annulus surrounding the fixation point from each resolution level.

4.3.3 View-Based Standardization. Recall from Section 2.2 that the principle of cortical magnification dictates that more cortical processing is devoted to regions closer to the fixation point than to more peripheral regions. This will be reflected in the model by converting the FOV into a polar representation (Figures 15 and 16) via the mapping

$$\begin{aligned} \hat{\mathcal{F}}(r, \theta) &= \mathcal{F}(\wp(r \cdot cos\ \theta), \wp(r \cdot sin\ \theta))\ , \quad \theta \in \{\Delta_\theta, 2 \cdot \Delta_\theta, \cdots, 360\} \\ r &= \sqrt{x^2 + y^2}\ , \qquad\qquad x, y \in \mathcal{V} \end{aligned} \tag{6}$$

where $\wp(\cdot)$ indicates the operation of rounding to the nearest integer, Δ_θ is equal to 360 divided by the desired angular granularity, and $\mathcal{V}$ is the collection of pixel locations making up the field of view. This process is accomplished for both $\mathcal{F}_{Mag}$ and $\mathcal{F}_{Orient}$. Every row in the resulting transformed views correspond directly to a radial in the associated circular FOV, with movement vertically in the polar view translating directly to rotation in the pre-transformed view.

One would next like to represent these FOVs in a framework such that rotational orientation of the circular FOV (and ultimately the sequence of FOVs) is irrelevant to the search/recognition task. This is accomplished by rotating (shifting vertically in polar coordinates) the views such that all elements are presented ***with respect to the orientation of the edge at the fixation point,*** ϕ_c. Note that the edge and bar magnitude maps both are shifted using the orientation of the *edge* at the fixation point. Additionally, all elements

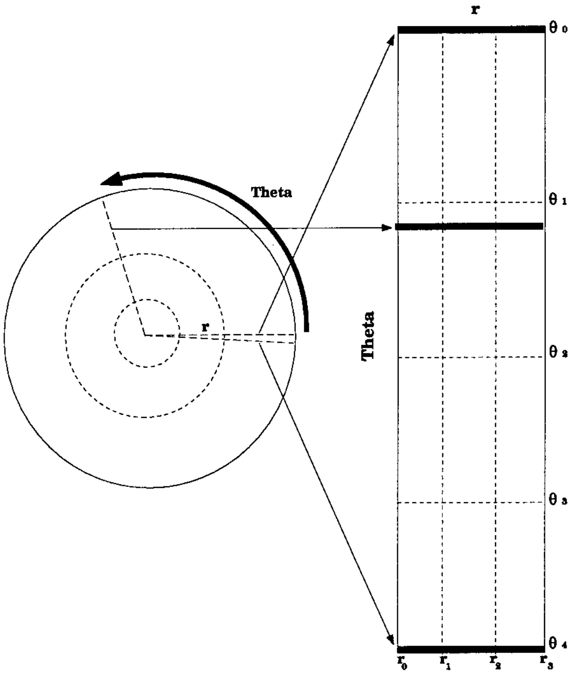


Figure 15. Transformation of FOV into polar space.

in the orientation maps $\mathcal{F}_{Orient}$ are transformed to provide orientation information with respect to the center edge. The following equations define the two mappings:

$$\hat{\mathcal{F}}_{Mag}(r,\theta) \rightarrow \hat{\mathcal{F}}(r,(\theta-\phi_c) \bmod 360) \quad (7)$$

$$\hat{\mathcal{F}}_{Orient}(\phi) \rightarrow \hat{\mathcal{F}}_{Orient}((\phi-\phi_c) \bmod 360) \quad (8)$$

Scale invariance is also implemented. During construction of the visual field for memorization, the outer radii of the rings surrounding the fixation point are defined as $r \times 2^k$, where r is the radius of the inner-most ring (the center disk), and k is the resolution level (with $k = 0$ at the original resolution). During search/recognition, however, the value of r can be varied during construction of the FOV, allowing examination of the image at multiple scales. The tremendous number of parallel cortical connections in the human brain [82] suggests that implementing such an approach in parallel is reasonable. Because

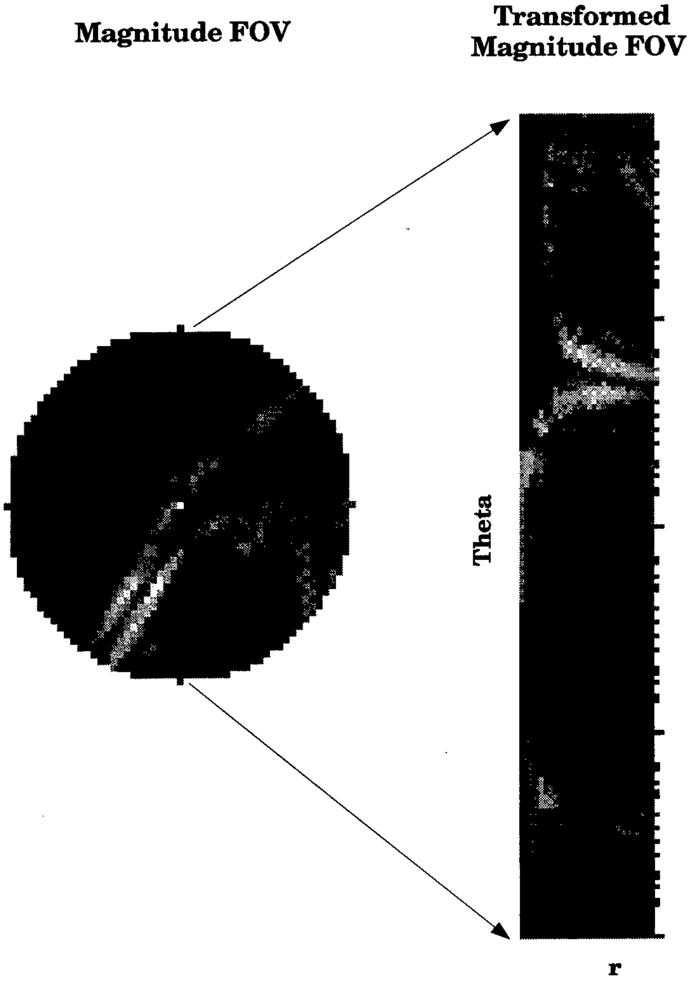


Figure 16. Polar transformation example.

the features used to describe the FOV are independent of its size, this approach works well.

4.3.4 Matrix Representation of the Field of View. Multiple histogram-based feature vectors are next extracted to represent the FOV. Because of the importance of maintaining information on the relationship between edges/bars in different portions of the visual field, SCAN-IT partitions it into eight regions: four regions representing quadrants of the circular visual field, three regions representing rings of increasing distance from the fixation point, and one region consisting of the entire visual field. The first four regions P_i

will contain the edges E determined by the relation

$$P_i = \{E_{R,\Theta} \mid r_0 \le R \le r_3,\ (\theta_i - \delta_\theta) \bmod 360 \le \Theta \le (\theta_{i+1} + \delta_\theta) \bmod 360\ \}, \quad i = 0, 1, 2, 3$$

The δ_θ term is provided to allow a limited overlap of each region with its two adjoining regions. This overlap increases the system's robustness to small rotational variations. The next three regions are defined in a similar manner by

$$P_j = \{E_{R,\Theta} \mid r_j \le R \le r_{j+1},\ \theta_0 \le \Theta \le \theta_4\}, \quad i = 0, 1, 2$$

As stated previously, the final region contains all elements within the visual field.

For each region, a normalized histogram of the magnitudes of the oriented edges is next constructed as

$$H_i(\alpha) = \frac{\sum\limits_{E_\alpha \in P_i} |E_\alpha|}{\sum\limits_{E \in P_i} |E|} \tag{9}$$

where $H_i(\alpha)$ is the normalized magnitude of the total response of edges at orientation α in region P_i, E_α is an edge with orientation α, and $|E_\alpha|$ is the raw magnitude of an edge at orientation α. The current FOV is thus memorized as a matrix whose rows represent the relative magnitude of a particular orientation in a particular region.

4.4 Saccadic Attraction

To determine a saccadic target, one must determine the point in the artificially constructed FOV that is the most interesting. In this case, the phrase interesting can be described in terms of edge intensities and distances from the current and past fixation points. We calculate the attraction of edge $E_{x,y}$ at spatial location x, y within the FOV as follows:

$$attraction(E_{x,y}) = [a \cdot |E_{x,y}| + b \cdot D_{x,y}] \cdot P_{x,y} \tag{10}$$

where $|E_{x,y}|$ is the magnitude of the response of edge $E_{x,y}$ at spatial location (x, y), $D_{x,y}$ is the distance of the edge from the fixation point, and $P_{x,y}$ is the penalty associated with the location of the edge due to a previous fixation at or near that point. a and b are experimentally tuned parameters used to adjust the relative importance of the edge magnitude and distance from the fixation point. For our trials (Chapter V), a value of $a = 1$ and $b = \frac{1.11}{r}$ worked well, where r is the radius of the circular field of view.

4.5 Memorization

Figure 17 shows a block diagram of SCAN-IT in memorization mode. An initial fixation point is first automatically selected by choosing the location with the greatest edge response at the lowest resolution, and the invariant FOV is constructed as described in Section 4.3.2. The feature matrix is then extracted and stored in what can be termed the 'what' memory. The attraction of each edge in the FOV is calculated, and the winning location is selected. The distance and angular direction (with respect to the orientation of the edge at the fixation point) to the winner is calculated and stored in 'where' memory, and the focus of attention is shifted to the new fixation point. This process is repeated for the desired number of saccades[2], and at its conclusion the object description has been stored as two matrices, one consisting of the concatenation of the feature matrices extracted from around each fixation, and one containing the angular and distance information needed to describe the path from one fixation to the next. Search mode can now be invoked to determine the presence of the memorized object in another image.

4.6 Search/Recognition

Figure 18 shows the model configuration during search and recognition. In the search task, a new image is presented to the model, and a saccadic scan of the image is performed. It is important to note that calculation of saccadic targets during this search process is performed identically as during memorization. One should expect, then, that if the

---

[2] Though selection of this number is somewhat ad hoc, it is primarily dependent on the size of the object to be memorized relative to the size of the visual FOV. Sufficient fixations must be made to memorize enough of the object to recognize it during a subsequent search.

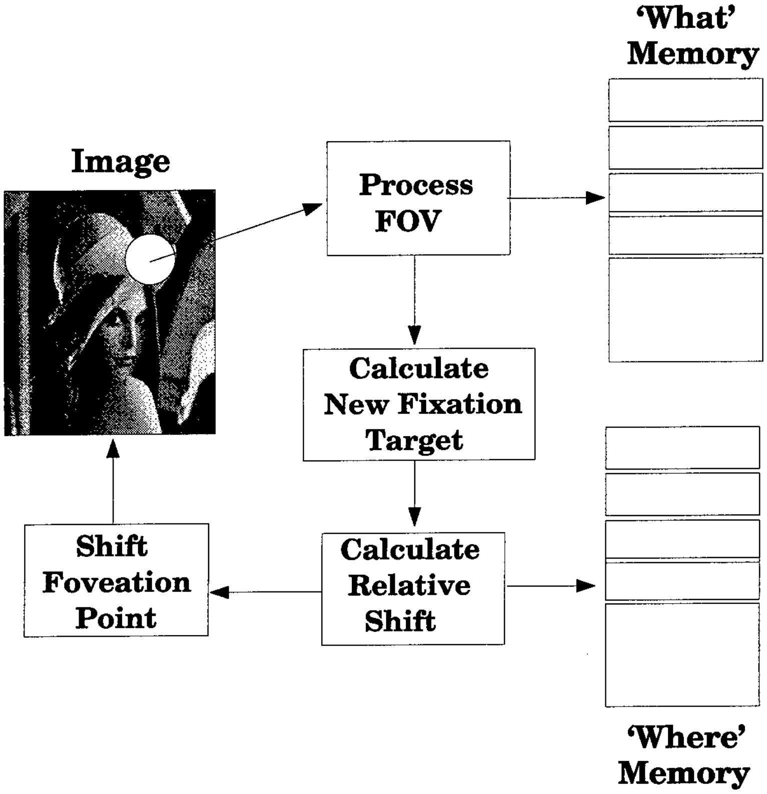


Figure 17. Vision Model in Memorization Mode.

memorized object/scene is present in the new image, a saccade will eventually produce a fixation at a point that was selected during memorization.

At each fixation, the visual field is again constructed and the edge features extracted, but rather than simply saving the data, this feature matrix is compared (via a Minkowski distance metric with parameter 2) with each of the feature matrices obtained during memorization, producing an error measure $\mathcal{E}$ for each matrix:

$$\mathcal{E}_m = \left(\sum_{(x,y)\in H_m} |H_{current}(x,y) - H_m(x,y)|^2 \right)^{\frac{1}{2}} \tag{11}$$

where $H_{current}$ is the feature matrix extracted from the current FOV, and H_m is the feature matrix extracted during the mth fixation of the memorization process.

If an error value for any of the memorized FOVs falls below a specified threshold, the hypothesis is formed that the object/scene is present in the new image. At this point

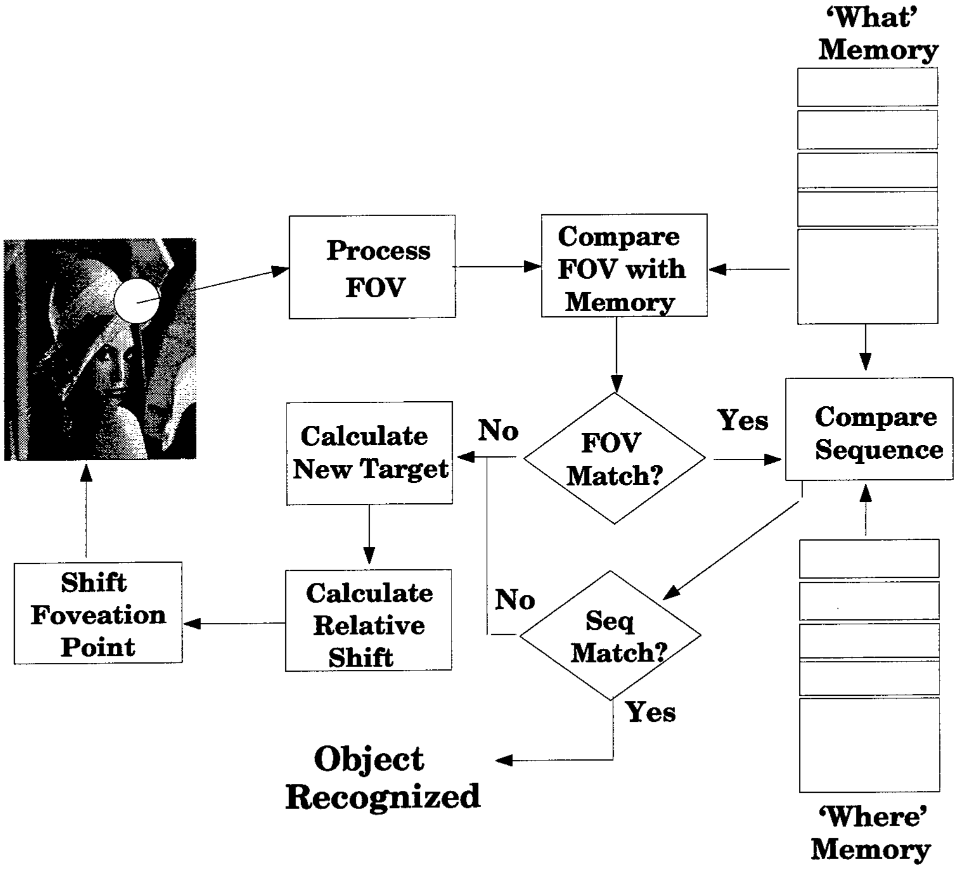


Figure 18. Vision Model in Search/Recognition Mode.

recognition mode is invoked, and the sequence of saccades recorded during memorization are calculated (with respect to the current fixation and center edge orientation) and repeated over the new image. At each fixation, the error between the current FOV and the corresponding memorized FOV is calculated, and if the mean of the errors is below a second error threshold, a new instance of the object/scene is recognized; if the expected features are not found, the model reverts to search mode. The second threshold is set higher than the first (for the trials reported in Chapter V, the second threshhold was set 20% higher than the first) to increase robustness to occlusions or distortions that may occur at fixations about the new image. If the image has been fully explored with no discovery of the object/scene, the search is terminated with the conclusion that the memorized object/scene is not present.

4.7 Summary

In this chapter SCAN-IT, a behavioral model of human vision was introduced and described in detail. In the next chapter, experimental trials are performed using SCAN-IT as a recognition engine.

V. Experimental Results using SCAN-IT

5.1 Introduction

The experimental trials were designed to demonstrate the utility of SCAN-IT in object memorization/recognition tasks. To do so, a building-block approach was used, starting with a relatively simple problem space, and advancing through increasingly complex realms. For each of the six sets of trials performed, a selected object was first memorized by SCAN-IT via a series of saccades. That object was then searched for within a target image, the characteristics of which could be affected by the following variations:

- Object Presence: The memorized object could be present or not in the target image.
- Rotation: If present, the memorized object could be subject to rotational variations.
- Scale: If present, the memorized object could be subject to scale variations.
- Background: The background of the target image could be benign (blank) or complex. If complex, the background could contain distractors spatially similar to the memorized object.

By creating target images with varying characteristics as just described, the following capabilities can then be demonstrated:

- Fundamental memorization, search, and recognition/rejection capabilities.
- Ability to correctly recognize/reject:
    - simple, segmented, black and white (B & W) objects in an image containing similar B & W objects
    - gray-scale objects in complex backgrounds not containing any objects spatially similar to the memorized object.
    - gray-scale objects in complex backgrounds containing objects spatially similar to the memorized object.
    - B & W objects in backgrounds containing a mixture of B & W and gray-scale objects

  - gray-scale objects in backgrounds containing a mixture of B & W and gray-scale objects.

- Ability to fuse individual recognition decisions to correctly recognize specified combinations of B & W and gray-scale objects.

Section 5.3 introduces extensions to SCAN-IT that will solve performance shortfalls found during the experimental trials. A context-based methodology will be applied to the recognition algorithm that will be seen to improve accuracy in the problem environments examined here.

5.2 Experimental Setup and Results

Six separate sets of experimental trials (for a total of 453 individual trials) were performed to address the capabilities listed in the previous section, with the goals and conditions of each set summarized in Table 1. The **Object** column specifies the object or class of object that was memorized for the particular set of trials, while the **Search Images** column lists the contents and type of target images to be searched.

One should note that two separate measures will be important in measuring performance in these memorization/recognition tasks: Sensitivity and Specificity. Sensitivity is defined as the True Positive rate, or the rate at which SCAN-IT correctly recognizes a memorized object when it is presented for search. Specificity, on the other hand, is defined as the True Negative rate, the rate at which an image not containing the memorized object is correctly rejected. Because both of these measures are important, it is appropriate that the test cases be balanced to provide an approximately equal number of trials to each of the object-present and object-absent classes. This was accomplished for the experiments presented here by putting all image variations (scale and rotation variants) containing the memorized object in one collection, and all target images not containing the object in another. An equal number of target images (or as close to equal as practical) were then randomly selected from each collection for testing. This balanced testing provides more meaningful results than simply testing all possible target combinations.

Table 1. Experimental goals and conditions

Trial	Purpose	Object	Search Images
1	Demonstrate fundamental memorization and recognition capability	B & W text (word)	B & W text (word) on blank backgrounds
2	Demonstrate capability to correctly recognize or reject B & W objects among similar B & W objects	B & W text (word)	B & W text (sentences) on blank background
3	Demonstrate capability to correctly recognize or reject gray-scale objects among dissimilar gray-scale objects	Gray-scale segmented faces	Gray-scale image from which face was segmented. No other faces in the image
4	Demonstrate capability to correctly recognize or reject gray-scale objects among similar gray-scale objects	Gray-scale segmented faces	Gray-scale image from which face was segmented. Other faces present in the image
5		Gray-scale cars	Gray-scale mosaic of 8 similar cars
6	Demonstrate capability to correctly recognize or reject B & W objects and/or gray-scale objects among mixed gray-scale and B & W objects	B & W text (words)	Image containing both B & W text and gray-scale objects
		Gray-scale objects	Image containing both B & W text and gray-scale objects

The target variants were produced by rotating and scaling the target images to produce a total of 8 variations of each image. Four rotational variations were used, 0, 30, 90, and 180 degrees, and each of those variations was subjected to two of three scale factors: 1 (for all trials), 1.5 (for the first four sets of trials), and 0.75 (for the last two sets of trials). As stated in the previous paragraph, target images were randomly drawn from these variants to avoid producing results showing bias toward (or against) any particular rotation/scale combination. Appendix A shows the objects and original (non-rotated, non-scaled) target images used for all trials reported on here. Selected examples of the target image variants are also provided in that appendix.

Because the final set of experimental trials performed here combines the functionalities demonstrated in each of the five previous sets, only the results from that experiment will be discussed in detail in the body of this document. Only the setup, results, and a brief discussion will be given for the first five trials.

5.2.1 Trial Set 1: Fundamental Capabilities. This case demonstrates that SCAN-IT can memorize, search for, and recognize simple objects under the most benign conditions. Black text on a white background provides for a simple object with no need for segmentation. Searching for this text in a target image that similarly consists of black text on a white background (where that text may or may not be the same as that memorized) should put little stress on the system, but will show basic capabilities in a quite straightforward way. Demonstrations of robustness to scale and rotational changes will also be shown here.

The text objects used for this set of trials were the following five words:

the rain in Spain falls

Each of these words was memorized in turn, and then tested against a selected sample from the collection of all words and their variants. To balance the testing, each word was tested against six variants of itself, and six variants of all the other words. This resulted in a total of 30 Sensitivity trials and 30 Specificity trials.

Figure 19a shows an example of one of the words to be memorized, Figures 19b and 19c show two of the target images containing the memorized word, and Figure 19d shows one of the non-object images. For each trial, the desired object was memorized via a sequence of four fixations, and SCAN-IT was allowed 25 saccades over each target image to determine the presence of the memorized word. Figure 20a shows the saccadic path obtained during object memorization, and Figure 20b shows the path taken during successful recognition of one of the target variants. The 25 fixations accomplished over an unfamiliar target (the word "rain") are shown in Figure 20c. Note that fixations have been made to points outside the visible text edges (points a biological system would not normally saccade to), but this is simply an artifact of the penalty region as described in Section 4.4. Recall that after every fixation, a penalty is applied to an area surrounding the fixation point to prevent too rapid a return to that area, thus permitting a more thorough exploration of the target image in a smaller amount of time. This penalty is allowed to

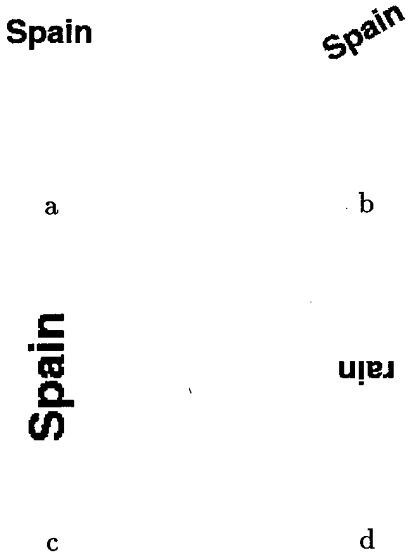
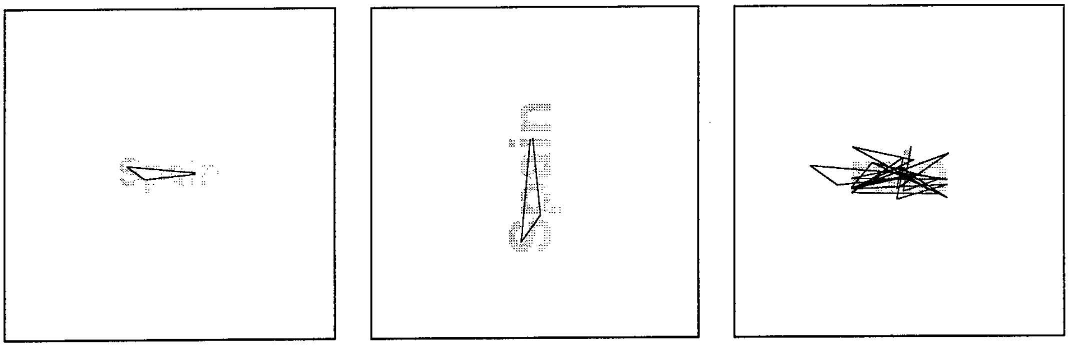


Figure 19. a) One of the words to be memorized, b) a rotational variant of the word, and c) a scale and rotational variant of the word. d) A rotational variant of a non-memorized word.

decrease over time (measured in number of saccades) to allow later exploration within that region. In the case of single word search as shown here, the total penalized region built over several saccades is large enough to virtually eliminate any edge attraction over the text itself, only leaving the small Gabor filter responses at distant locations to act as attractors. As can be seen, though, those saccades are uncommon, and only occur here because the image being searched is so small and simple in content. This effect was even more pronounced on the smaller words, but did not affect recognition results.

Table 2 shows the results from the 60 trials, with the following terms defining the performance measures used (the same terms will be used for all remaining trials as well):

- **True Positive (TP):** Target correctly identified as containing memorized object.
- **True Negative (TN):** Target correctly identified as not containing the memorized object.



a b c

Figure 20. a) Saccadic path obtained during object memorization. b) Saccadic path taken during successful recognition of target variant. c) Search path for the memorized object in a target image not containing the target

- **False Positive (FP):** Target incorrectly identified as containing the memorized object, when in fact it did not.
- **False Negative (FN):** Target incorrectly identified as not containing the memorized object, when in fact it did.
- **Sensitivity (SENS):** The True Positive rate.
- **Specificity (SPEC):** The True Negative rate.

Table 2. Trial 1 Results

TP	TN	FP	FN	SENS	SPEC
30/30	25/30	5/30	0/30	100 %	83.3 %

SCAN-IT was thus successful in recognizing every instance of the memorized word, but did not appear to be quite as successful in correctly rejecting non-memorized objects. This is somewhat misleading, however, as four of the five recognition 'errors' occurred as SCAN-IT was examining the words **rain** and **Spain** for the presence of the word **in**. Indeed, the last two letters of each target word were correctly recognized as matching the memorized word, leading to an actual specificity rate of 97 %.

This was a set of simple trials to show fundamental memorization and recognition capabilities, but occurred in a very simple problem domain (recognizing a single word in an image containing a single word). The next set of trials increases the complexity of the problem by increasing the size of the search space (introducing target images with multiple words).

5.2.2 Trial Set 2: Word Recognition in Sentences. For these trials, the same words were memorized as in the first, but the search complexity was increased by using complete sentences as the target images. The process of detecting a specific word in a collection of text is sometimes known as word spotting or keyword spotting [36,47].

Figure 21 shows the three sentences used for these trials.

the rain in Spain
falls mainly
in the plains

the Maine rain
falls plainly
in east Wayne

This weather
can grate on
one's nerves

Figure 21. The three sentences serving as the base target images for the second set of trials.

Note that the words **the, rain, in, Spain** and **falls** are all present in Sentence 1 (i.e., all of the memorized words are present), **the, rain, in,** and **falls** are present in Sentence 2 (four of the memorized words are present), and none of the memorized words are present in Sentence 3. By creating the target variants as described previously, sufficient target images were randomly selected to test 31 object-present cases and 29 object-absent cases. SCAN-IT was allowed 30 saccades to determine the presence of the memorized word in the sentence. Figure 22 shows an example of one of the memorized words, a succesful recognition of the word, and the search path by SCAN-IT over a target image that did not contain the word. The dashed lines in Figure 22b and c indicate the search path prior to recognition (where in 22c the memorized object was not found). The results from these trials are listed in Table 3.

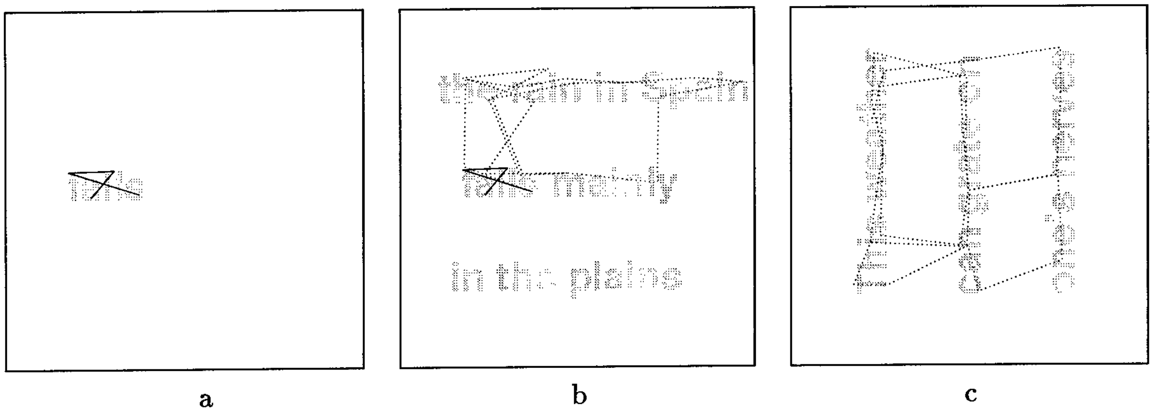


Figure 22. a) Memorization scan of the word **falls**. b) Successful recognition (dashed line indicates the search path performed by SCAN-IT). c) Search path for the word **falls** in a sentence not containing the word

Table 3. Trial 2 Results

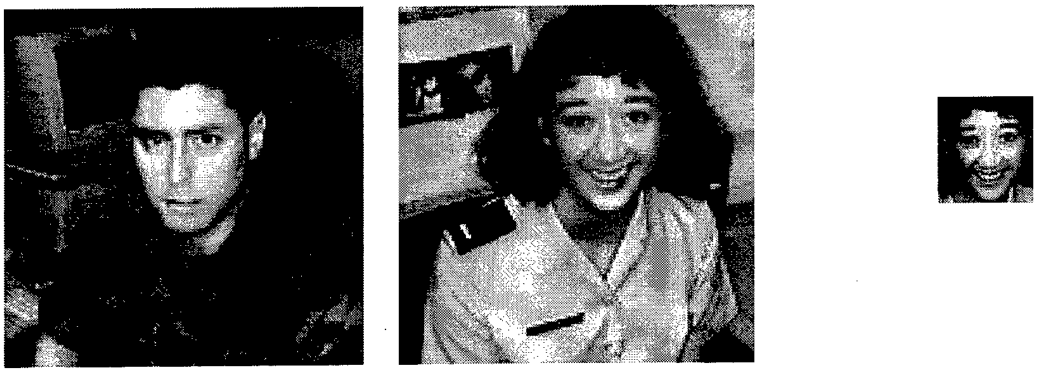
TP	TN	FP	FN	SENS	SPEC
28/31	23/29	6/29	3/31	90.3 %	79.3 %

SCAN-IT was generally successful in recognizing instances of the memorized objects, but again the Specificity results were not quite as positive. Analysis of the data revealed that two of the six False Positives, as with the cases in the first set of trials, were actually detection of the memorized word as part of another word (in this case, **the** was found within the word **weather**). A more accurate Specificity value to report would probably be 85.2%. Two more of the False Positives resulted from the detection of the word **rain** in the **rate** portion of **grate**. Other failures were attributed to either the limited number of saccades permitted for object detection or to a detection threshhold not optimized to this problem set (indeed, the same detection threshhold was used throughout the testing to demonstrate the utility of SCAN-IT in an arbitrary environment). Section 5.3 will propose a solution to the threshholding problem and the saccade limitation.

The next step in the evaluation process was to test SCAN-IT against more interesting, gray-scale objects and target images.

5.2.3 Trial Set 3: Memorization/Recognition of Faces in Complex Backgrounds.

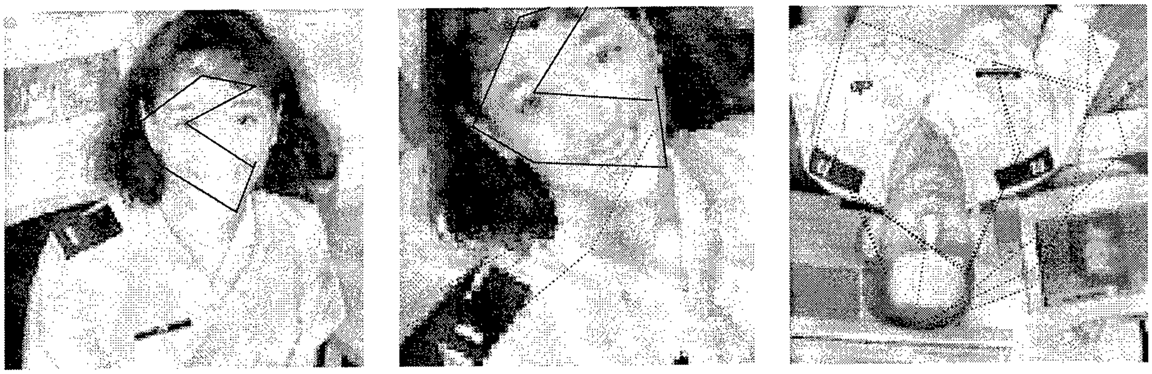
For these trials, five faces were extracted from gray-scale images to serve as the memorized objects. After memorization of each segmented face via eight fixations, SCAN-IT was allowed to scan the selected target variants for 30 saccades to determine its presence. Figure 23 shows two of the five images from which the faces were extracted, as well as one of those segmented faces.



a b c

Figure 23. a) and b) Two of the images from which the faces for Trial Set 3 were extracted. c) One of the extracted faces to be memorized.

Figure 24 shows examples of the memorization scan over a face, a successful recognition of that face, and a 30-saccade scan of an image not containing the memorized face. The final results from the 60 trials performed are provided in Table 4.



a b c

Figure 24. a) Memorization scan of one of the faces used for Trial Set 3. b) Successful recognition of the memorized face. c) Search of a target image not containing the memorized face.

Table 4. Trial 3 Results

TP	**TN**	**FP**	**FN**	**SENS**	**SPEC**
30/30	30/30	0/30	0/30	100 %	100 %

Note that for the first time a perfect recognition score was obtained. This would seem counterintuitive at first, as the images used for these trials are of higher complexity than those containing only simple black and white letters. However, the added complexity is actually one of the major reasons for the superior performance. Recall that specificity was the problem in the first two trials; that is, mistakes were made in believing a word was present that actually was not. But those B & W word objects, consisting primarily of interconnected straight lines and predictable curved lines, do not have nearly the random variability found in a natural image as used for these trials. Thus the discriminability between objects in gray-scale, natural images is likely to be higher than in a structured, B & W environment, a conclusion that certainly appears to be supported here. In addition, the segmented faces memorized here are larger (in proportion of the total image size) than the words used previously, allowing more information, and more useful information, to be memorized.

The next step, then, is to evaluate the performance of SCAN-IT in a gray-scale environment again, but with a more difficult target. In the next set of trials, faces will again be memorized, but the spatial area of those faces will take up a smaller proportion of the total image, and additional, possibly similar, faces will be included in the target images to act as distractors.

5.2.4 Trial Set 4: Memorization/Recognition of Faces in Complex Backgrounds Containing other Faces. In the previous set of trials, the face extracted from each image was the only object in the image with face-like qualities. For the trials in this fourth set, faces were extracted from images containing multiple faces, with the goal of determining SCAN-IT's ability to ignore distractors similar to the memorized object. The original **Gatlin** image from the Matlab© development environment was divided into four different images, each containing three faces (the original image and two of those sub-images are shown in Figure 25). Individual faces were then extracted from these sub-images, memorized (via a sequence of five fixations), and searched for among the target variants. Because of the increased complexity of these target images, the number of saccades SCAN-IT was allowed to accomplish in its search was increased to 60, but no other parametric changes were made.

Figure 26 shows examples of a memorized face, a successful recognition of that face, and a search of a target image not containing the face. The results from the fourth set of trials are presented in Table 5

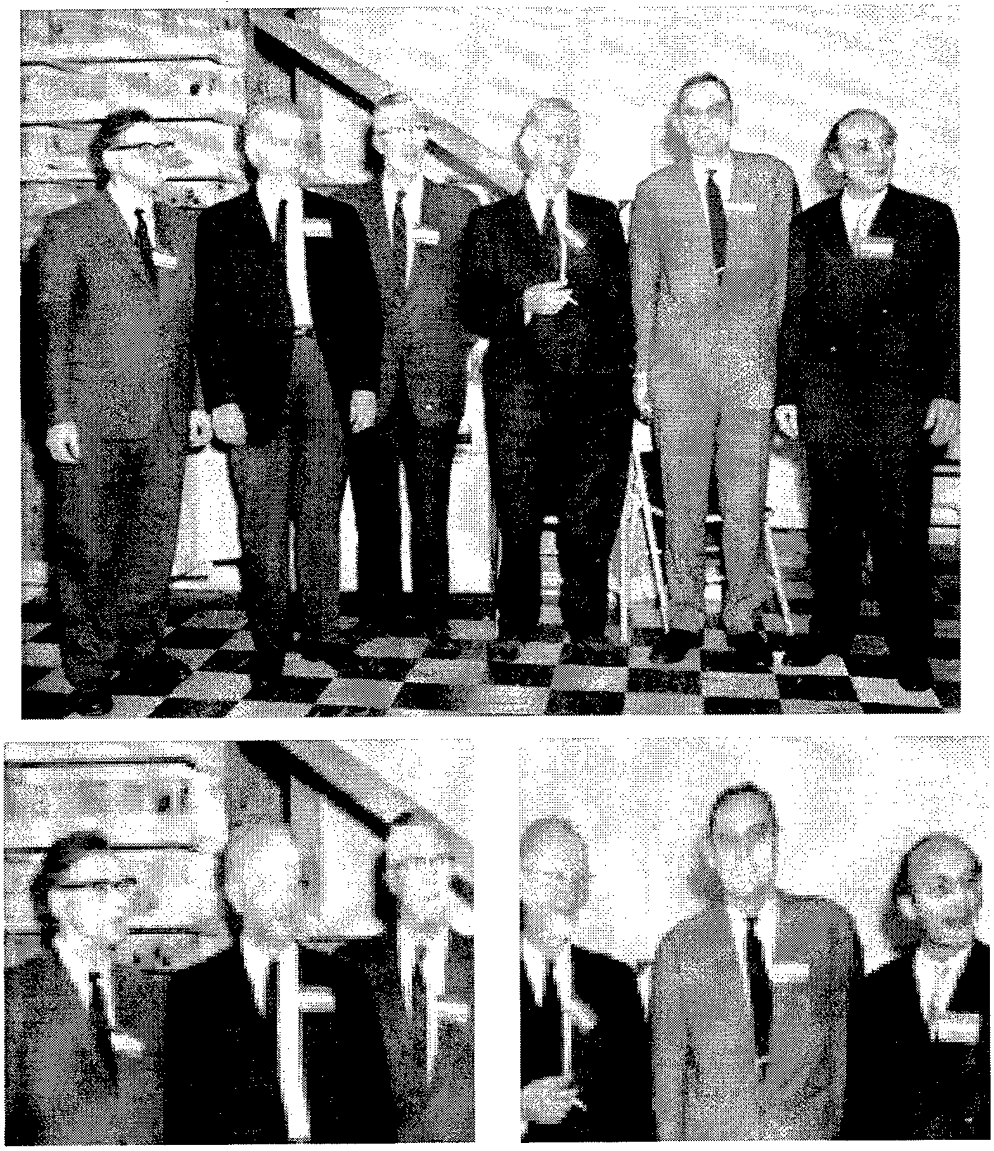


Figure 25. The original Gatlin image and two of the six sub-images used for Trial Set 4 (from the Matlab© development environment).

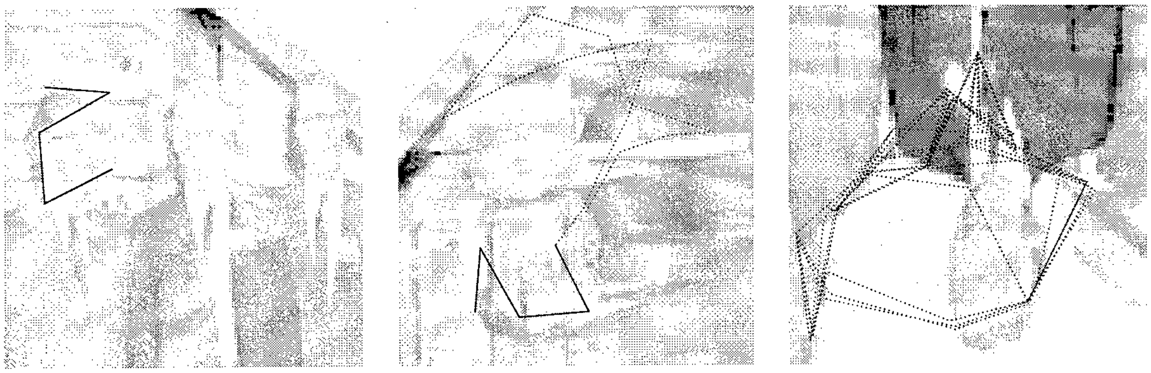


Figure 26. a) Memorization scan of one of the faces used for Trial Set 3. b) Successful recognition of the memorized face. c) Search of a target image not containing the memorized face.

Table 5. Trial 4 Results

TP	**TN**	**FP**	**FN**	**SENS**	**SPEC**
29/33	32/32	0/30	4/30	87.9 %	100 %

Specificity was again perfect for this trial set, but Sensitivity decreased from the prior trial. Analyzing the results showed that of the four False Negatives, two appeared to be caused by a large portion of the face being cropped by the 30 degree rotation process (because the target variants were randomly chosen, unless a majority of the face was cut off, the image remained eligible for testing). In the other two cases, even after 60 saccades the memorized face had been saccaded to only a few times, and never to memorized fixation points, so no recognition was made. Because the system was designed to fully explore an image *given a sufficient number of saccades*, confidence remains high that those faces would have been detected if the saccade number had been set higher [1]. However, limitations on the number of saccades necessary to discover a memorized object is an interesting and important quantity, so it was decided not to perform any optimization just to improve the results. This issue will be further addressed in Section 5.3

---

[1] To provide anecdotal evidence that this is indeed true, SCAN-IT was allowed to continue the search process on one of the False Negative items, and at the 75th fixation the face was properly recognized.

The next set of trials moves to even more complex gray-scale objects, cars set among multiple similar car distractors.

5.2.5 Trial Set 5: Memorization/Recognition of cars in Complex Backgrounds Containing other Cars. For the next increase in target image complexity, a collection of individual cars from the Corel© Image CD collection formed the collection from which all objects were selected. A total of 13 cars were used for these trials, of which five were randomly selected to to be memorized. Ten image mosaics, each consisting of eight of the 13 cars, were then created in such a manner that each memorized car was found in exactly five of the mosaics. The order of the individual cars within each mosaic was also randomly determined, and to form the complete collection of target images, the same four rotational variants were used as in the previous trials, but the scale factor used was 0.75 instead of 1.5. In other words, the scaled images for these trials (and all subsequent trials) were smaller than the originals. Figure 27 shows two of the individual cars used and two of the mosaics constructed for testing. A total of 100 trials were performed using these mosaics, with

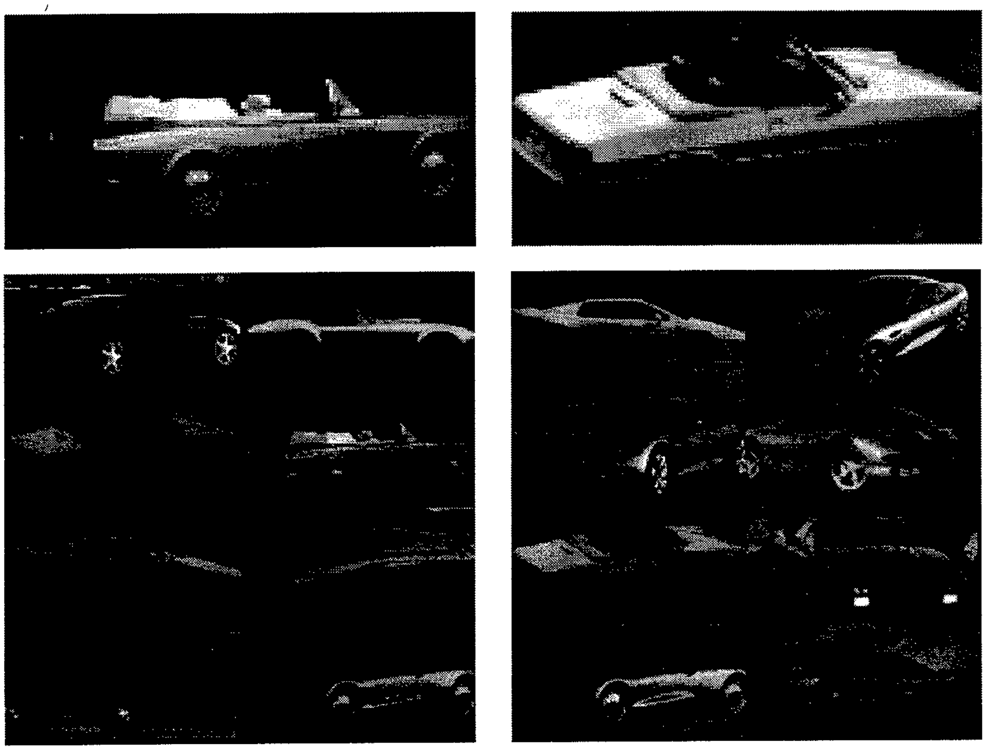


Figure 27. a) Two of the individual cars and two of the car mosaics used for Trial Set 5.

individual cars memorized via a 5-fixation sequence and SCAN-IT permitted 60 saccades to locate a memorized car in a mosaic. Figure 28 shows examples of a memorized car, a successful recognition of that car, and a search of a target mosaic not containing the car. The results from the trials are shown in Table 6.

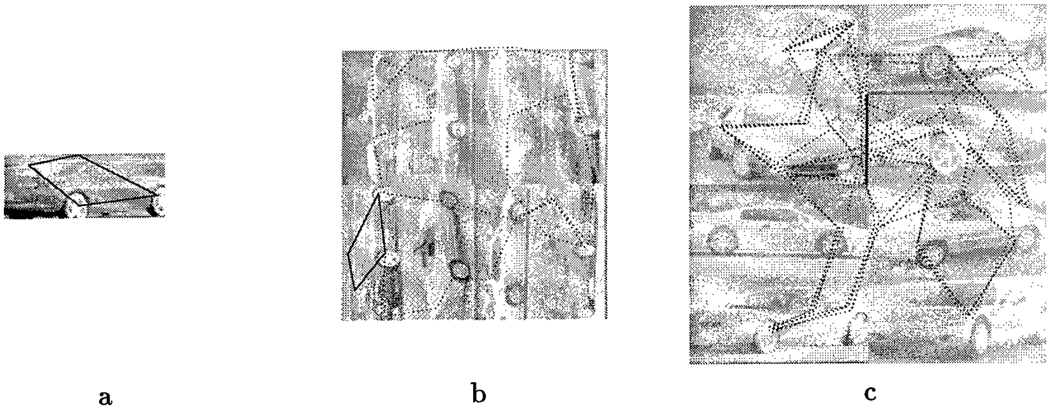


Figure 28. a) Memorization scan of one of the cars used for Trial Set 5. b) Successful recognition of the memorized car. c) Search of a target mosaic not containing the memorized car.

Table 6. Trial 5 Results

TP	**TN**	**FP**	**FN**	**SENS**	**SPEC**
45/50	49/50	1/50	5/50	90 %	98 %

As in the previous set of trials, excellent Specificity was obtained, but Sensitivity seemed not quite as robust. Data analysis showed that four of the five False Negatives were obtained when testing just one of the memorized cars against the target variants. In each of the failures with that particular car, points on the car were fixated at most twice, and in two cases the car was never fixated on during the search process. Thus, it is again believed that increasing the number of saccades permitted would have increased the performance, but not doing so provides more insight into potential limitations of SCAN-IT.

In the next section, a detailed discussion is provided on the conduct of and results from trial set number six.

5.2.6 Trial Set 6: Memorization/Recognition of Text and/or Cars in Target Images Comprised of Text and Cars. The final set of experimental trials builds on the functionalities demonstrated in the five previous sets to provide a comprehensive demonstration of SCAN-IT's ability to memorize and recognize arbitrary text/car objects in arbitrary target scenes. Additionally, a simple fusion scheme will be implemented that will solve the problem of concurrently recognizing multiple objects within a single target image.

Objects to be memorized included both text (white text on a dark background) and the same cars as used in the previous set of trials, and the target images were each a combination of text and one of five of the cars. The final goal of these trials then will be to correctly recognize a combination of specified text as well as a specified car. Figure 29 shows two of the ten target images used.

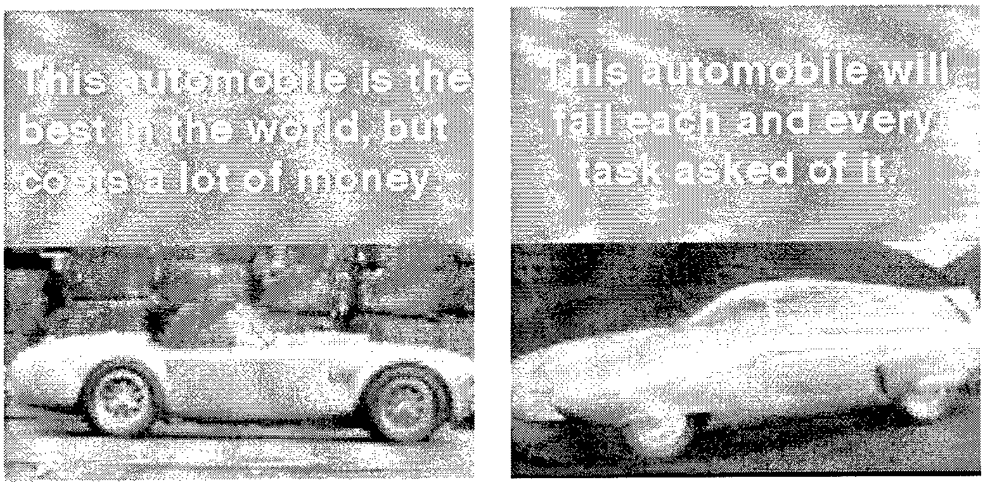

This automobile is the best in the world, but costs a lot of money.
This automobile will fail each and every task asked of it.

Figure 29. Two of the target images used for Trial Set 6.

The group of objects to be memorized consisted of the five different cars and the four textual phrases

"best in the world"
"costs a lot"
"fail each and every"
"task asked of it"

The first two phrases are from the sentence "This is the best automobile in the world, but costs a lot of money." The second two are from the sentence "This automobile will fail each and every task asked of it."

It is interesting at this point to observe exactly what is being examined during the memorization process. Figure 30 shows the portions of one of the cars memorized after one fixation, three fixations, and all eight fixations. This is a typical result from the memorization process, and one notes that at the end of the memorization process a vast majority of the object of interest has been examined. This is the ideal result, and should lead to the best recognition performance with minimum false alarms (False Positives).

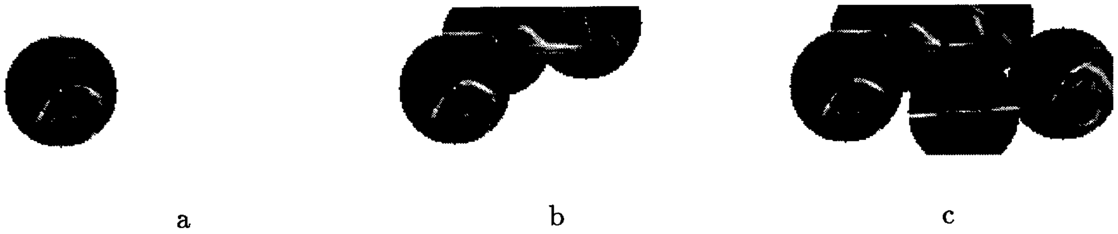


Figure 30. Portions of a car examined during the memorization process after a) 1, b) 3, and c) 8 saccades.

108 experimental trials were conducted against the target images (where for each trial, the target image was randomly selected from the pool of 80 available variants), with each object memorized via eight fixations, and SCAN-IT allowed 60 saccades to find the object if present. Figure 31 shows examples of memorized objects, successful recognition of those objects, and search of target images not containing the objects. The top row of figures shows testing with a car object, and the bottom with a text object.

Processing the individual car/text searches in parallel allows for the identification of target images that contain both a specified text object and a specified car object. One might imagine desiring the selection of images containing a picture of one's favorite car and a textual caption describing some quality of the car. 20 different combinations of text and car existed (four text objects and 5 car objects), and each of those combinations was tested against. To balance the testing, 20 Specificity cases were also tested, where the specific text/car combination was not present in the target image. For this experiment, the fused recognition decision was a simple logical ANDing of the individual object recognition decisions.

Results from the individual object trials are given in Table 7. Table 8 gives the results from the fused decision testing.

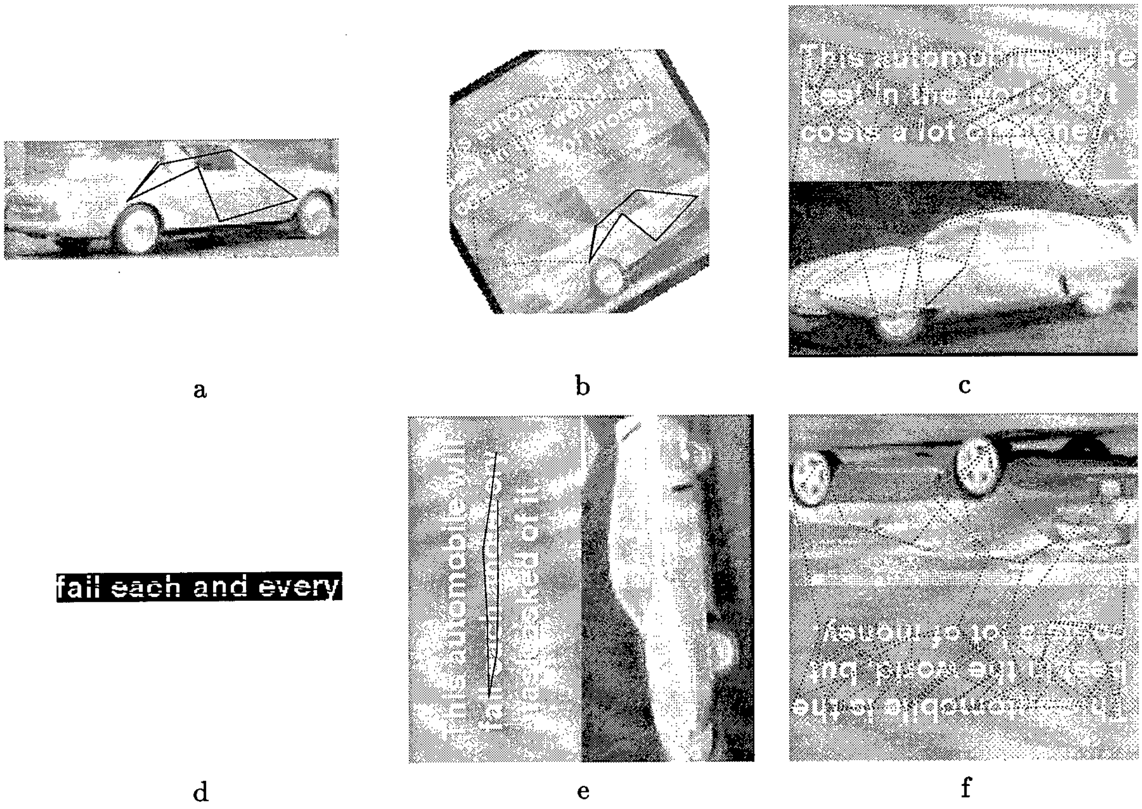


Figure 31. a) and d) Memorization scan of one of the objects used for Trial Set 6. b) and e) Successful recognition of the memorized object. c) and f) Search of a target image not containing the memorized object.

Table 7. Trial 6 Results

TP	TN	FP	FN	SENS	SPEC
53/54	51/54	3/54	1/54	98 %	94.4 %

Results from the memorization and search/recognition of either textual or gray-scale objects show that SCAN-IT is indeed flexible enough to search arbitrary scenes for these objects. Additionally, the perfect fused recognition decisions show that SCAN-IT provides

Table 8. Trial 6 Fusion Results

Fused TP	**Fused TN**	**Fused FP**	**Fused FN**	**SENS**	**SPEC**
20/20	20/20	0/20	0/20	100 %	100 %

sufficient information to allow the combination of individual decisions to select target images containing multiple desired objects.

the system is quite capable of combining individual decisions to select target images that contain multiple desired objects.

Both Sensitivity and Specificity for the individual cases are quite good, but it would be instructive to determine the reason for the recognition failures that did occur. The three False Positives resulted while textual phrases were being sought in the target images; Figure 32 shows the search scanpaths and false recognition in each of those cases. The top row shows the mis-identification on the target image and the bottom row shows the original memorization of the word object.

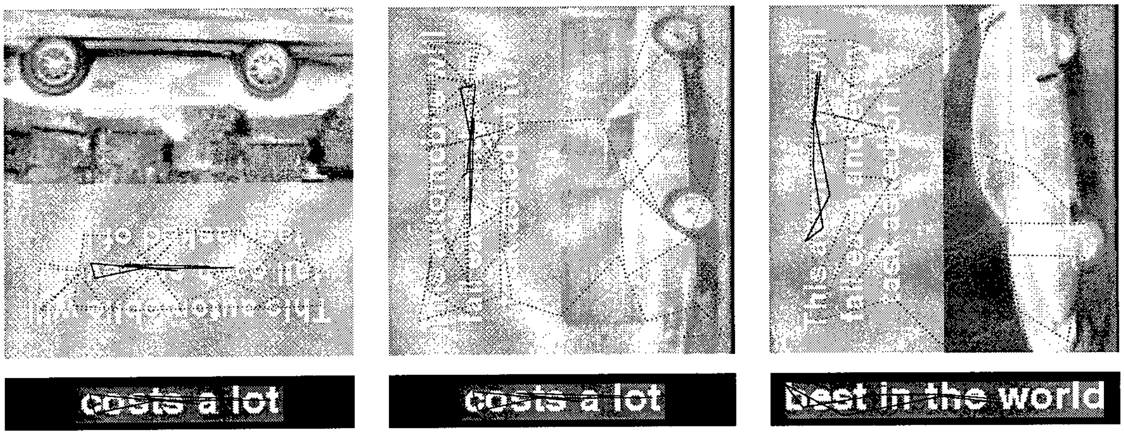
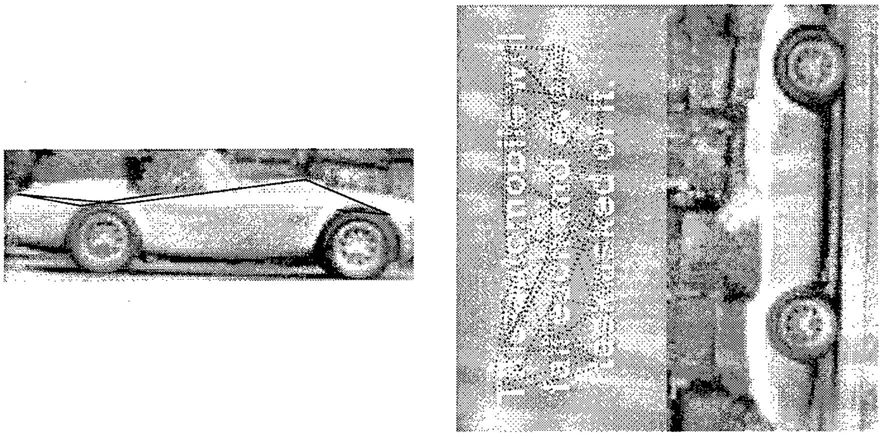


Figure 32. False Positive results from Trial Set 6. The top row shows the search path and false recognitions, and the bottom row shows the original memorization of the object after 8 foveations.

Two of the errors resulted while searching for the phrase **costs a lot** (Figures 32a and b), in which SCAN-IT found the object at the same location in the same sentence on two different target images. The last occurred when the phrase **best in the world** was mistakenly found in the same sentence. It is interesting that the specificity errors all occurred while searches for text object were being performed. While the similarity of the recognized text to the memorized phrases is not immediately apparent to the naked eye, it can be argued that the same explanation provided in the section describing Trial Set number three applies here. That is, the restricted nature of the text (straight lines, curves with constant radius, etc) results in less discriminability than is normally found in natural images. Indeed, post-experimental analysis of the data revealed that in each of these False Positive cases, the recognition error (the Minkowski distance between the memorized object and the recognized object) was higher (about 10 % on average) than in any of the cases in which an object was correctly recognized. This means that the recognition threshhold of SCAN-IT could have been tuned to provide perfect Specificity results on this set of trials. However, it is quite often more important to not miss true occurrences of a desired target than it is to incorrectly identify the target as being present (that is, True Positives are often more important than False Positives). For example, a bank using an automated system to provide services may prefer to have a higher False Positive rate in order to avoid wrongly refusing service and alienating customers. SCAN-IT's threshhold is set to meet that condition here, but a solution to the threshholding problem is proposed in Section 5.3 that will produce recognition decisions based on contextual information in the target image.

Figure 33a shows the memorized scanpath of the car object for the one False Negative case, while 33b shows the search scanpath accomplished during that trial. The fundamental reason for non-recognition in this case is obvious: a fixation was never made to the memorized car in the target image. This is an issue that can be addressed in several ways. First, because SCAN-IT is guaranteed to fully explore any image given sufficient time (sufficient number of saccades), the number of search saccades allowed could be increased (as seen in Sec 5.2.4, allowing an increased number of search saccades indeed resulted in correct recognition of a memorized object not previously found). Section 5.3 will address this issue by proposing extensions to SCAN-IT that will avoid the whole issue of specifying saccade limits.



a b

Figure 33. False Negative result from Trial Set 6. a) shows the memorization of the car, and b) shows the search scan path.

The parameters of the inhibition imposed on the region surrounding a fixation could also be adjusted to alter the frequency with which an area is revisited. It was noticable in all of the experimental trials conducted over the combination text/image targets that most fixations were made to the text portions, and generally only after these regions were well explored did the focus of attention move to the car. This is a direct result of the stronger edge gradients (and hence stronger attraction) produced by the B & W text. By increasing the size of the penalized region and/or by increasing the amount of time it takes the penalized region to recover to its original values, fixations will tend to move to the car object after fewer saccades. Once SCAN-IT is given the opportunity to directly explore the memorized car, it should be rapidly recognized, as is supported by all the results presented here.

However, SCAN-IT was designed to perform in an arbitrary environment, and the parameters applied are thus somewhat of a compromise between working best for simple B & W text and for complex gray-scale imagery. Methods do exist to detect text within imagery, generally based on textural and stroke cues (see, for example, [47, 76, 80]), and if one knows in advance that the object being searched for is text, then SCAN-IT could be modified to use these techniques. This, though, would detract from the generalization currently achieved. With the present parameters, excellent Sensitivity is already achieved for both types of memorized objects, so additional optimization is considered neither necessary nor desirable to increase accuracy on this particular (randomly selected) test set.

The following section provides a tabular summary of the results from all the experimental trials conducted.

5.2.7 Summary of Experimental Results. Table 9 consolidates the results from all six experimental trials of SCAN-IT. As discussed in the previous sections, for different types of memorized objects and target images Sensitivity or Specificity results may have been slightly better, but over all 453 trials the two performance measures are quite closely matched. The successful performance of SCAN-IT under experimentally controlled conditions has been shown, but it has also been pointed out that some shortfalls may exist in the parameterization balance between good Sensitivity and good Specificity performance in the general case. Thus, in the next section a technique is introduced that will obviate the need to spend time determining the proper recognition threshhold for a given problem set.

Table 9. Summary of Experimental Results

Trial Set	**TP**	**TN**	**FP**	**FN**	**SENS**	**SPEC**
1	30/30	25/30	5/30	0/30	100 %	83.3 %
2	28/31	23/29	6/29	3/31	90.3 %	79.3 %
3	30/30	30/30	0/30	0/30	100 %	100 %
4	29/33	32/32	0/30	4/30	87.9 %	100 %
5	45/50	49/50	1/50	5/50	90 %	98 %
6	53/54	51/54	3/54	1/54	98 %	94.4 %
Total	215/228	210/225	15/225	13/228	94.0 %	93.0 %

5.3 Using Context to Solve the Sensitivity-vs-Specificity Problem

The problem of proper threshholding is a common one in pattern recognition tasks [15, 58, 65, 73]. Determination of the discriminant boundary that will appropriately separate multiple classes such that both Sensitivity and Specificity values meet the desired (and contradictory) goals is not always possible. What this means is class discriminability is almost always a compromise, where a balance is reached between correct classification and correct rejection. This concept can be illustrated by the simple example shown in Figure 34. Given an ensemble of data belonging to either class *x* or class *o* (as determined

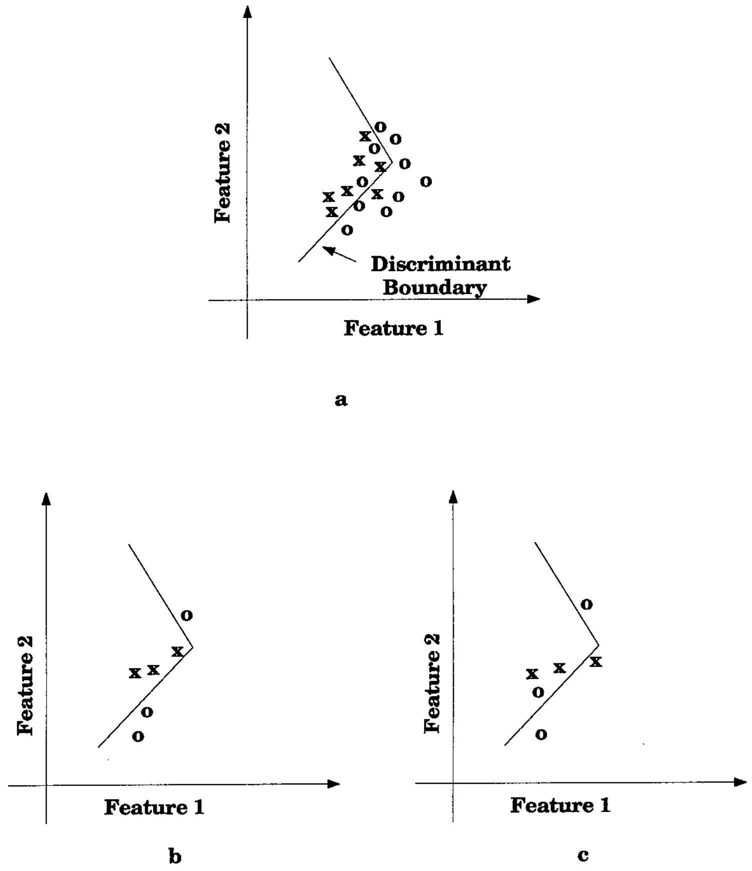


Figure 34. a) Discriminant boundary constructed to separate two classes of data to some precision determined by the designer. b) Example of data points perfectly separated by the boundary. c) Example of data points producing False Positives.

by a function of Feature 1 and Feature 2), some constructive technique can be used to build a discriminant boundary that will separate data of those classes to some arbitrary precision. In Figure 34a, a simple, piecewise-linear discriminant has been constructed that will separate the data to a precision that is satisfactory to the designer of the classifier. Given the data points shown in 34b, classification performance would be perfect, with the discriminant boundary perfectly separating the two classes. Figure 34c, however, shows data points that would not be properly classified by the classifier. A data point that in actuality belongs to class x, but lies to the right of the discriminant boundary, will be declared to belong to class o, generating a False Positive if one is searching for objects of class o. Likewise, any data points belonging to class o but lying to the left of the boundary will also generate False Positives when searching for objects of class x. Thus, construction of the discriminant boundary is important, but given that it can be virtually impossible to make it perfect (especially in the case of incompletely representative data, which is usually the case), often ends up being a compromise between producing no False Positives and producing some acceptable rate of False Positives.

In the case of SCAN-IT, the compromise is imposed primarily by the recognition threshhold, which should allow correct recognition of memorized objects while at the same time rejecting images that do not contain the object (i.e., separate the classes of object-present and object-absent). As seen in the experimental trials, there are cases where Specificity has been negatively affected (False Positives have occurred because of a too-high recognition threshhold), and in the following paragraphs a solution is proposed that will be seen to remedy the problem. This solution will be fully generalizable to a wide variety of pattern recognition problems and will not adversely affect the Sensitivity of the classification system.

The general idea is to use image context to determine whether or not a declared recognition is correct. To collect this contextual information, the operation of SCAN-IT can be slightly altered. The upper limit on number of search saccades allowed will be removed, and the decision of object-present or object-absent will no longer be made by simply achieving a Minkowski distance below the recognition threshhold. Instead, when that threshhold is breached (for the remainder of this section, the term *possible detection* or simply *detection* will be used to indicate this situation), the distance value will be recorded, but no recognition decision will be made; SCAN-IT will be allowed to continue searching.

One should note at this point that a context-based approach will not be amenable to all problems. In cases where time (measured in actual time or in CPU cycles) is at a premium, it may be more desirable to use SCAN-IT as originally designed, establish as good a recognition threshhold as one is able (perhaps based on training cases with sample objects and sample target images), and declare successful recognition when that threshhold is breached. However, if the problem environment is such that more time is available to undertake additional processing, the contextual recognition described in the following paragraphs will improve the recognition results. In other words, the application of SCAN-IT can be completely generalizable based on the time restrictions under which one is attempting to solve a problem.

After some user-specified number of possible detections[2] (instances of possible object recognition), the search process will be stopped and the resulting error (Minkowski distance) data will be examined. Recall that a fundamental assumption behind SCAN-IT's functionality is that features generated from a memorized object will be substantially different from those generated by any other object; i.e., the error will be small when a memorized object is scanned, and large when any other object is scanned. If the error values recorded over the possible detections are closely clustered (the variance is small), probability is high that a correct instance of recognition did *not* occur during any of the detections. These error values, then, can be thought of as describing the memorized object within the context of *non*-object space within the target image. If, however, there is a high variance among the values, where one or more of the errors are substantially lower in value than the rest, probability is high that those smaller errors correspond to instances of correct recognition.

To test this hypothesis, the collection of a balanced set of True Positive and False Positive decisions was desired. Three FPs were generated during the final set of experimental trials, but to provide additional data points, the recognition threshhold was raised and the trials repeated, generating more FPs. Because the text objects (as opposed to the car objects) presented the most Specificity problems, only data obtained testing with memorized text objects was used here. For each memorized word, two target images were selected that had led to FP results with the non-modified implementation of SCAN-IT (including the three cases that had occurred in the final trials), and two that had led to

---

[2] That number will primarily depend on the number of fixations with which the original object was memorized; more of these detections should be allowed than there were fixations, to ensure some saccades are made to regions not originally memorized

TP results; a total of 16 targets, then, was to be tested against with the extended version of SCAN-IT.

After modifying SCAN-IT as described above, eleven possible detections were recorded from each of the object-absent cases, and eleven from the randomly selected set of object-present cases. If the contextually-based discrimination concept is sound, one would expect the detections for the object-absent cases to have tightly clustered error values. For the object-present cases, on the other hand, one would expect to find one or more error values that are significantly less than the others (which will correspond to the instances of correct recognition – True Positives).

Figure 35 shows that this is indeed the case. Data for each row corresponds to exactly one of the text phrases tested against, with the plot on the left showing results from testing against the object-absent cases (the memorized object was *not* in the target image), and the plot on the right showing results from testing against the object-present cases (the memorized object *was* present). Each plot shows the error values (Minkowski distances, vertical axis) obtained at each of the eleven possible detections for two separate trials, for a total of 16 trials shown in the eight plots. It is apparent that for the object-present trials (right plots), there is a much greater variability than for the object-absent trials, reflecting the fact that both true and false detections were recorded.

Though the variability in recognition distances is evident in the plots, a quantitative method is still desired to determine which particular recognition instances will be considered correct. Recall that the intent of this context-based decision scheme is to reduce the importance of the recognition threshhold, to instead rely on the relative distances between errors measured during multiple possible detections. But in the object-present cases, there is still no a priori knowledge of how many of those detections will be correct (or incorrect), so the obvious statistical measure of variance may not be the ideal one. Instead, a metric will be developed based on a different assumption: that error values close to the smallest error are likely to indicate an instance of possible object recognition. To define what close means, examine the relative distances between the maximum (D_{max}) and minimum (D_{min}) distance values recorded for a particular object over the 11 recognition instances,

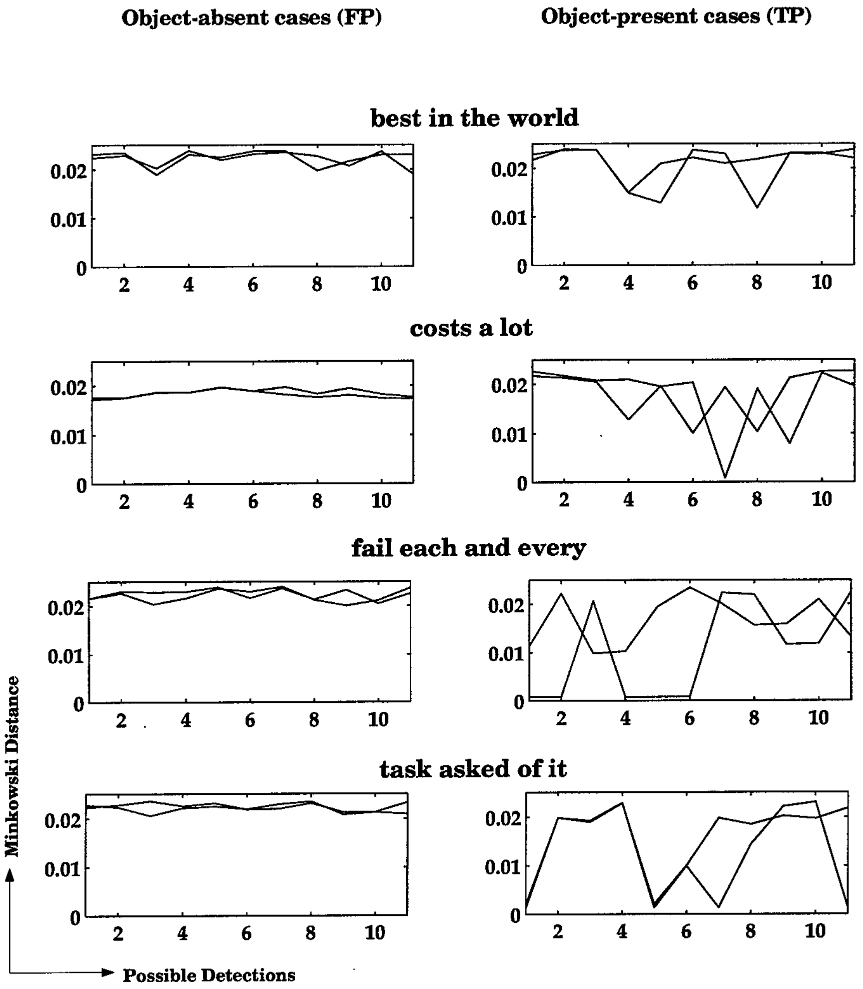


Figure 35. Plots of Minkowski distances calculated for each recognition instance reported by SCAN-IT. Each row corresponds to one of the four text objects, for which two object-absent and two object-present cases were tested (the left and right plots respectively).

and select a threshhold T midway between the two:

$$T = D_{min} + \frac{D_{max} - D_{min}}{2} \tag{12}$$

Thus, any detection producing errors less than or equal to T are assumed to be ***possibly*** correct, and any producing errors greater than T are assumed to be false. The statistical means of error values associated with each of these two classes can now be calculated as $\hat{D}_{max}$ and $\hat{D}_{min}$, as can the distance $\Delta\hat{D}$ between those two means for each memorized object:

$$\hat{D}_{min} = \frac{1}{m}\sum_{a\in A} a \quad , A = \{D|D \leq T\},\ m = \text{size}(A) \tag{13}$$

$$\hat{D}_{max} = \frac{1}{n}\sum_{b\in B} b \quad , B = \{D|D > T\},\ n = \text{size}(B) \tag{14}$$

$$\Delta\hat{D} = \hat{D}_{max} - \hat{D}_{min} \tag{15}$$

In the equations above, m refers to the number of detections with error values less than or equal to T, and n to the number with error values greater than T. A simple graphical example of the calculation of these values is shown in Figure 36. Given the six error values (1, 3, 12, 13, 14, and 17) derived from six detections of a single object, D_{max} and D_{min} are determined, T is calculated per Equation 12, and $\hat{D}_{min}$, $\hat{D}_{max}$, and $\Delta\hat{D}$ are calculated from Equations 13, 14, and 15.

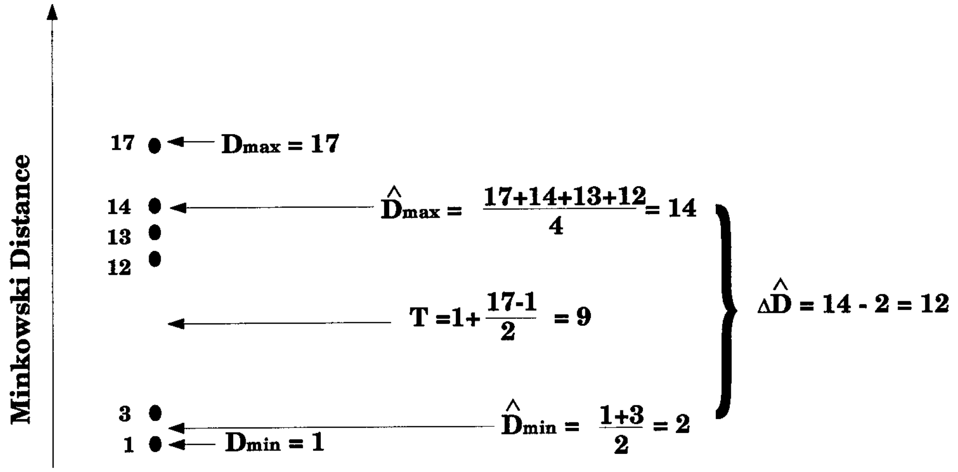


Figure 36. Graphical example of calculation of $\hat{D}_{max}$, $\hat{D}_{min}$, and $\Delta\hat{D}$ for context-based recognition

Figure 37 shows $\Delta\hat{D}$ for the eight object-present cases (solid line) and the eight object-absent cases (dashed line). As can be seen, there is generally better than an order of magnitude greater distance between the two classes (no-object and possible-object) for the object-present cases than for the object-absent cases. Thus $\Delta\hat{D}$ serves as an excellent metric to decide whether a true recognition has in fact occurred (among the multiple

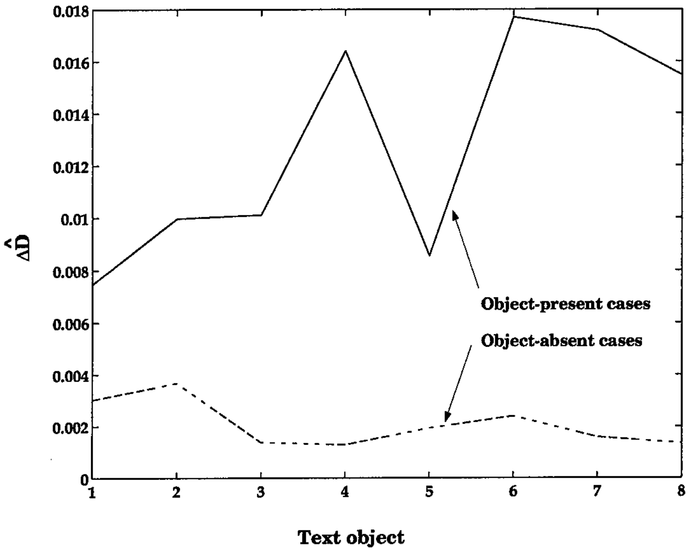


Figure 37. Comparison of $\Delta\hat{D}$ for the eight object-present cases (solid line) and the eight object-absent cases (dashed line).

possible detections reported by SCAN-IT). The accuracy of the recognition threshhold is no longer of vital importance as long as it is between the minimum for the object-present cases and the maximum for the object-absent cases. Once the decision is made that at least one of the possible detections was correct, those detections belonging to the possible-object class (that is, those producing errors less than T) can be rank-ordered by the confidence that each detections is a True Positive. The detection producing the lowest error will produce the highest confidence, and that confidence will decrease as the error increases. One simple confidence measure that could be used is

$$C_i = \frac{D_{min}}{D_i} \tag{16}$$

where C_i is the confidence that detection i is a True Positive, D_{min} is the minimum Minkowski distance produced by any of the detections, and D_i is the Minkowski distance produced by detection i.

To summarize and recap the extensions to SCAN-IT: first, the limit on maximum number of search saccades allowed is removed. The recognition scheme is altered such that when the recognition threshhold is breached, the Minkowski distance is recorded, but the search is allowed to continue. Some set number of possible detections will thus be accomplished, with no decision yet on whether those detections are correct or not. Once those detections are accomplished, the error (distance) values will be examined, and a measure $\Delta\hat{D}$ will be calculated that will provide information on the relative distance between the cluster centers of non-recognition and possible-recognition instances. If $\Delta\hat{D}$ is greater than a loosely set threshhold, the conclusion is reached that the possible-recognition instances are actually each object-present cases with associated confidence based on the corresponding Minkowski distance.

Recall that in the final set of experimental trials (text or gray-scale objects sought in combined text and gray-scale target images), three False Positives were produced. Implementing the modified SCAN-IT as described above, those trials were re-run, and using a final context threshhold of 0.006 (based on the values shown in Figure 37, all three cases were determined to in fact *not* produce any correct recognitions ($\Delta\hat{D}$ values for the three cases were 0.0037, 0.0014, and 0.0013). In other words, Specificity for the trials was raised from 94.4 % to 100 % (Sensitivity remained unaffected at 100 %. Note that though a threshhold was still required, the value in this case could be much more loosely established because of the relatively large delta between $\Delta\hat{D}$ for targets in the object-present class and $\Delta\hat{D}$ for targets in the object-absent classes. Though only anecdotal results are presented here, it is firmly believed that this characteristic will be found in any problem set. The concept of context-based recognition, then, can be generalized to any given problem environment.

Though the context-based recognition scheme proposed here will improve Specificity performance of SCAN-IT, there is still a price to be paid. As mentioned in the beginning of the discussion, using this methodology will generally require a substantial increase in time and CPU activity. In environments that rely primarily on timeliness, using the original, non-extended SCAN-IT may be more appropriate. When time is not as critical, and when accuracy of results is more important, the extension of SCAN-IT into a contextual realm will be highly desirable. Additionally, as computers get faster and as the recognition algorithm is made more efficient, the cost will go down, making use of the extended SCAN-IT more palatable in any problem environment.

In the next chapter, SCAN-IT is applied to a real-world problem, retrieval of specified video imagery from within a video sequence.

VI. Video Retrieval: A Real-world SCAN-IT Demonstration

In the previous chapter, SCAN-IT's memorization, search, and recognition capabilities were shown experimentally, but it is also constructive to see its performance when applied to a real-world problem. The problem addressed here will be in the area of image retrieval, where SCAN-IT will be used to retrieve frames containing specified objects from video imagery.

6.1 Introduction to Image Retrieval

Living in the age of information has resulted in the generation and availability of vast stores of digital data. But though this data may be available, it is becoming more and more evident that tools for finding data of interest to a particular user, and ignoring irrelevant data, must be developed. In the case of textual data, search mechanisms already exist that can find specific tokens (words and phrases, for example), as well as relationships between tokens within a document (see [25, 26]).

A related task is the retrieval of digital imagery from large image databases (including video sequences), but search methods here are not nearly as straightforward or robust as for modern textual searches. Though textual tags can be linked with an image, the probability of a user selecting the proper words or word-combinations to retrieve every image of interest is small. A typical human being, on the other hand, has little or no problem recognizing images containing instances of some particular object of interest (for example, "Find every image containing an airplane."), and in fact seems to perform this task with ease. We appear to have an inherent capability to quickly and effortlessly find things that "look like . . ."

Much research has been accomplished in the area of content-based image retrieval, with a variety of techniques used in the attempt to find the right images based on some measure of correctness or similarity. Most of these techniques involve extracting visual characteristics of objects within the image (e.g., texture [35, 38, 39, 45, 46, 49, 54], shape [1, 22, 51, 52, 54, 69], and color [7, 23, 38, 39, 54, 56, 71]) and using various similarity metrics to compare sample images with a desired image.

Few retrieval methods, however, are based on biological systems, which is where SCAN-IT can play a role. Because SCAN-IT was designed to emulate (in some sense) certain characteristics of the human visual system, and because it has been shown to perform well in recognizing specified objects in arbitrary images, it is potentially well suited to image retrieval tasks. In the next section, the specific problem will be introduced, and SCAN-IT will be shown to be capable of solving the problem.

6.2 The Problem: Finding Specified Video Imagery

The tremendous quantities of video imagery available today (as well as the projected growth of production of such imagery) can cause problems if one desires access to only one small portion of it. Manually viewing the imagery in the hope of eventually finding the portion of interest is inefficient to the extreme, so a method is needed to automate that process. SCAN-IT will be used to provide that automation.

One method that can be used to improve search efficiency is to convert a sequence of video imagery into a much smaller set of images that still capture the contents of the sequence. Such an idea was explored in the course of this research, where SCAN-IT was used to locate fiducial markers in one video frame, recognize the same markers in another frame, and use that data to co-register the frames. This process was accomplished for a sequence of frames, and then the contents of all frames were fused into a single output mosaic based on contrast characteristics of the human visual system. Details of that work can be found in Appendix B. Another approach to the video retrieval problem is to examine individual frames for the presence of some specified object or objects. That is the approach to be discussed here in this section.

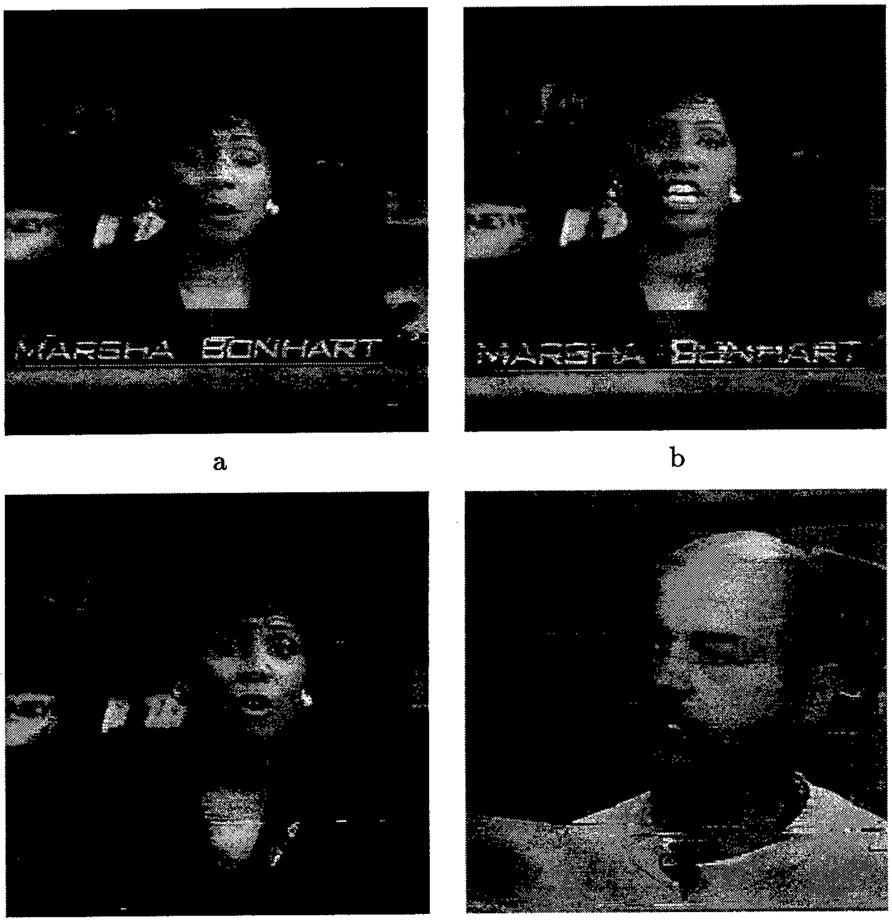
The specific problem domain to be examined here is a video sequence of a newscaster reporting a story. The goal will be to extract video frames from the sequence that contain the reporter, where the evidence of the reporter's presence will be given by both the face (face object) AND a caption containing the individual's last name (text object) [1]. Note that the requirement is not necessarily to locate every frame that contains those two

---

[1] Methods to extract text from within images do exist (for example, [47, 76, 80] but a single generalized tool for arbitrary text *or* objects is not currently available

attributes, but to locate at least one that does. Such a video marker is meant to lead the user to the proper section of the overall sequence.

Examples of the frames used for this demonstration are shown in Figure 38, where 38a shows the first frame in the sequence, 38b and 38c shows two intermediate frames, and 38d shows the last frame. The entire video sequence was comprised of 425 frames, recorded at 15 frames per second, for a total length of just under 30 seconds. The sequence was broken up into 15 sub-sequences, each representing about a two second span of time. A sample frame was selected from each sub-sequence to test against, making for a more efficient search process. The assumption in this case, of course, was that the reporter and

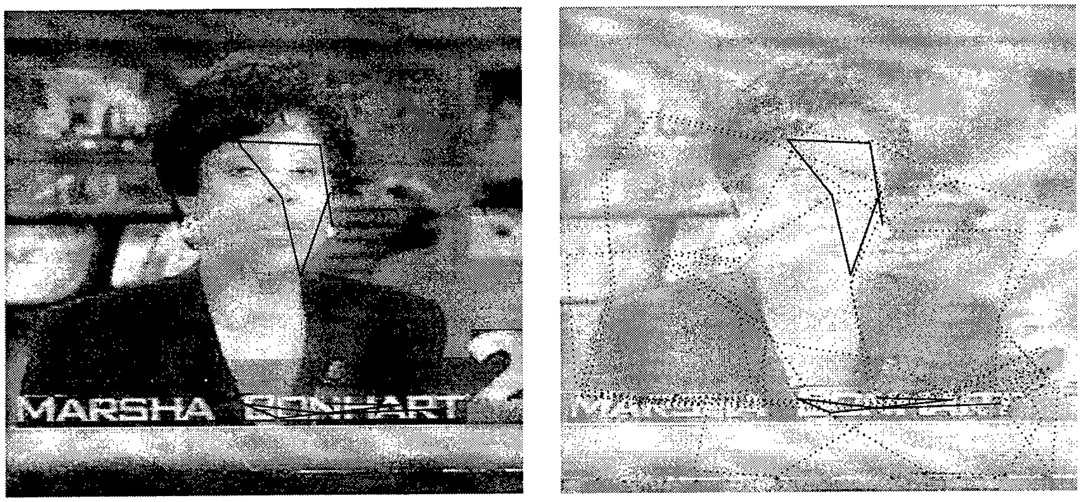


a b c d

Figure 38. Frames from the video sequence used for demonstrating SCAN-IT. a) The first frame in the sequence. b) and c) Two of the intermediate frames. d) The last frame of the sequence.

the caption will be on-screen for at least four seconds, so the two-second sampling should capture at least one of those frames.

The face and the caption from frame 5 were first memorized by SCAN-IT via six fixations (with the memorization scans shown in Figure 39a), and then the 15 target frames were selected for search. Contextually-based recognition as described in the previous chapter was applied for these cases, with 11 possible detections permitted for each target image, but no limitation on total number of search saccades. The distance metric $\Delta\hat{D}$ was calculated as described in Section 5.3 for each object/frame combination (i.e., the memorized face versus each frame and the memorized text versus each frame), and the same loose threshhold applied as given there. The result was one frame meeting the object-present criteria for both face and text. The search for, and subsequent recognition of each object in that frame is shown in Figure 39b.



a b

Figure 39. a) Memorization of the face and last name of the reporter, each via a sequence of six scans. b) Subsequent recognition of the face and last name in a different frame.

Subsequent analysis of the data showed that ten of the selected frames actually contained the reporter's face, seven of which were appropriately recognized by SCAN-IT (see Figure 40), and two of the frames contained the text, one of which was selected by SCAN-IT. All other frames were appropriately rejected. It should be noted that in

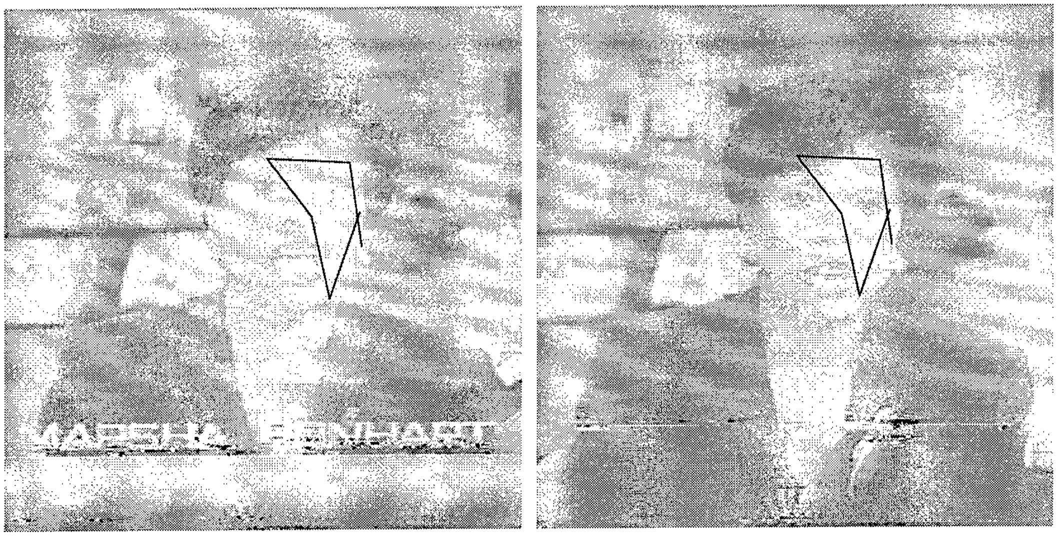


Figure 40. Correct recognition of the face object in Frames 40 and 87.

this demonstration, the memorized face-object was not identical to the face-objects in the frames being searched. This is shown clearly in Figure 40, where the reporter has very different facial expressions (from each other and from the memorized face shown in Figure 39a, but was still correctly recognized by SCAN-IT.

6.3 Conclusion

This successful demonstration of SCAN-IT's recognition capabilities support the utility of the model in a real-world environment. It should be stressed that this was purely an anecdotal demonstration, without the accompanying tabular data or in-depth analyses produced for the experimental trials of Chapter V. Instead, this was intended to give the reader a better feel for the manner in which SCAN-IT can be applied to actual, messy problems.

The next chapter will provide a summary of the research accomplished for this dissertation and will review the contributions made.

VII. Conclusion

7.1 Research Contributions

The primary objective of this research was to develop a new computational vision model based on physiological and behavioral characteristics of the human visual system. Such a model should be capable of detecting, identifying, and verifying the presence of arbitrary objects within arbitrary visual scenes and video sequences. The objective was successfully met with the development of SCAN-IT, Saccadic Control for Active Neural Identification of Targets. The results from this research can be summarized by the following contributions:

1. A new model of the human visual tract was introduced that emulates the function of biological neurons in the Retina, Lateral Geniculate Nucleus, and Primary Visual Cortex.
2. A view-based standardization process was developed that simulates the physiological characteristics of contrast sensitivity, contrast normalization, and cortical magnification to project an arbitrary visual field of view into a standardized, simulated cortical map.
3. A scanning algorithm was developed that in some sense emulates the saccadic behavior of the human visual system, permitting complete exploration of arbitrary visual scenes.
4. The characteristics in 1-3 above were synergistically combined to produce a recognition model that was shown to be capable of recognizing arbitrary objects within arbitrary scenes.
5. A context-based recognition paradigm was introduced that solves the general thresholding problem encountered in pattern recognition processes.

7.2 Future Directions

The contributions outlined above provide a firm foundation for future work on physiologically-motivated object detection. Though SCAN-IT has been shown to be effective at memorization and recognition, areas exist which could be improved to increase performance, efficiency, and utility.

- The capability to automatically adapt parameters to the type of object being memorized, as well as the type of target image being searched, would provide for a more efficient search and better Sensitivity and Specificity performance. For example, it was shown in the course of this research that SCAN-IT performed differently for text searches than it did for gray-scale object searches; the fixed parameters were established such that any arbitrary search produced acceptable performance.
- Recognition capability will benefit from the development of a single-number quality metric that will define the similarity of any given detection of an object to the memorized object. The confidence measure produced by the context-based recognition methodology introduced in this research is a step toward that capability, but determination of a context-threshold, though more easily established than the recognition threshold, is still required.
- Saccadic activity is an area that merits further exploration. SCAN-IT uses a very simple attraction scheme to determine the fixation sequence, a scheme based primarily on the gradient intensity of edges within the Field of View. No attempt was made here to perform saccadic activity in the same manner as biological systems; thorough exploration of the visual scene was determined to be of greater value, and this goal could be accomplished with the relatively basic method employed here. However, research shows that humans are more efficient at exploring scenes, using a sophisticated guidance and control system to direct the saccadic activity to portions of the scene that contain information that will more readily allow memorization and recognition. It will be worthwhile to explore introduction of a similar capability into SCAN-IT.
- Extension of SCAN-IT to permit recognition of memorized objects in different poses (vice different scales and orientations). The current saccadic scanning method performs well if spatial relationships (between edges and bars) in the target object are very similar to those in the memorized object. However, large changes in the pose of

the target object will adversely affect the recognition capabilities by changing those relationships, resulting in SCAN-IT not looking at the correct location to extract features. The capability to adapt the saccadic activity to the image being examined will solve that problem.

Appendix A. Experimental Trial Data

This appendix contains images of the object data used during the experimental trials detailed in Chapter V. Each of the different types of objects is shown in original scale and orientation, and examples of the objects in scaled and/or rotated configuration are included as well.

A.1 Trial Set 1 Data

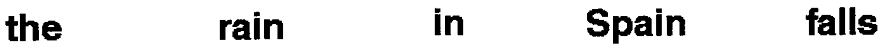


Figure 41. The five objects memorized for Trial Set 1. These objects also served as target images.

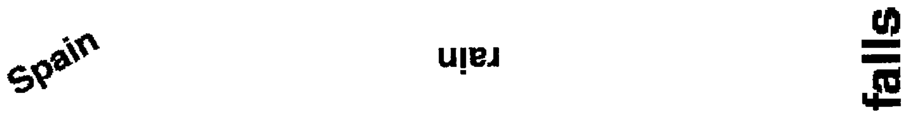


Figure 42. Examples of the target image variants used for Trial Set 1.

A.2 Trial Set 2 Data

For the second set of experimental trials, the objects to be memorized were the same as used during the first set, and are shown in Figure 41.

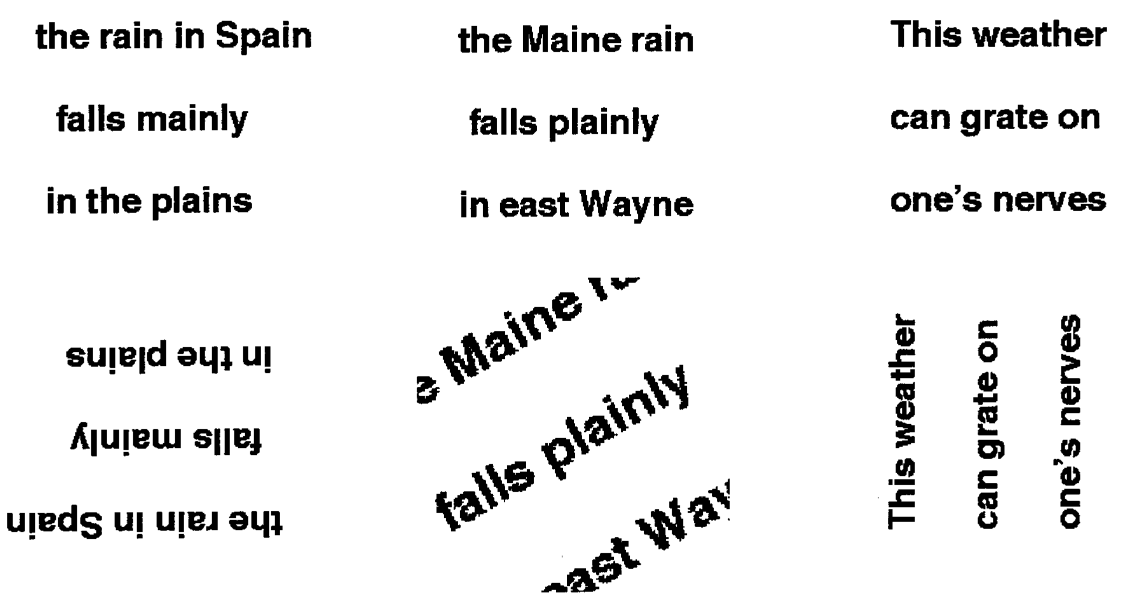


Figure 43. The original three target images (non-rotated/scaled) tested against for Trial Set 2 (top row) and examples of three of the variants (bottom row).

A.3 Trial Set 3 Data

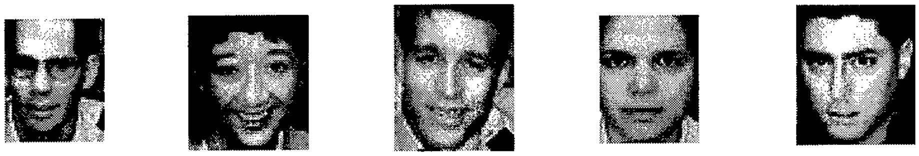


Figure 44. The five face objects memorized for Trial Set 3.

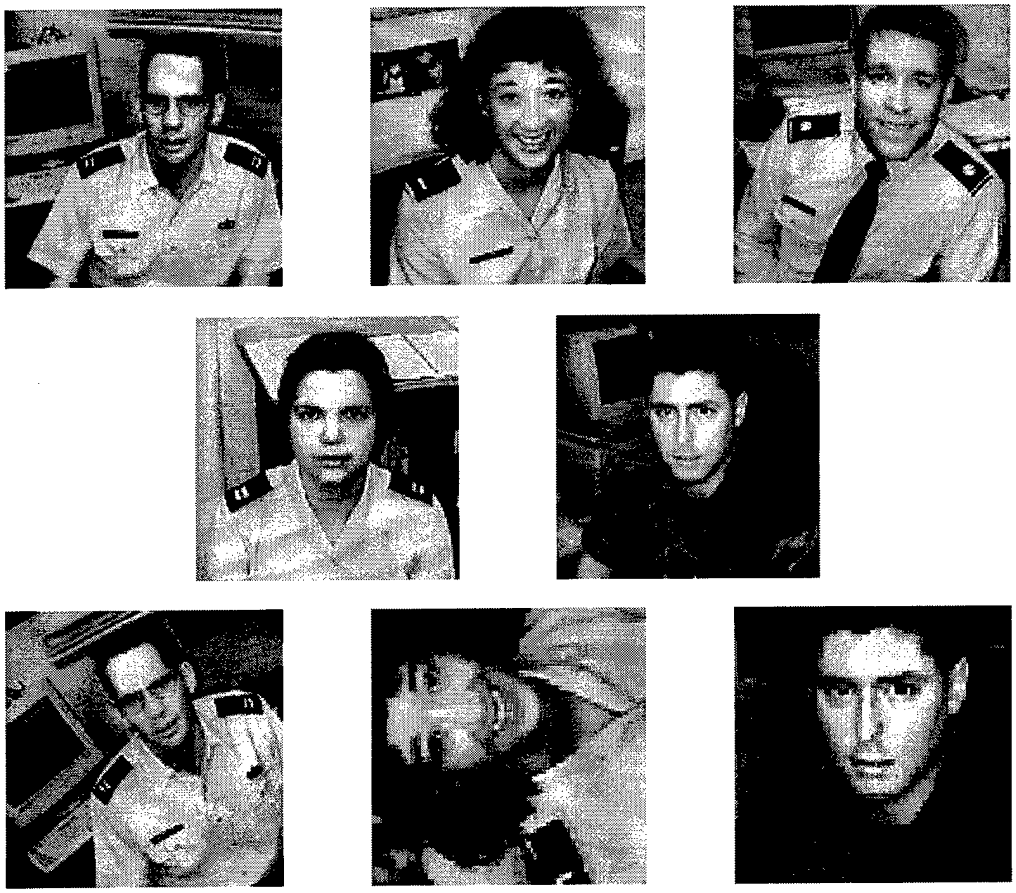


Figure 45. The original five target images tested against for Trial Set 3 (top two rows) and examples of three of the variants (bottom row).

A.4 Trial Set 4 Data

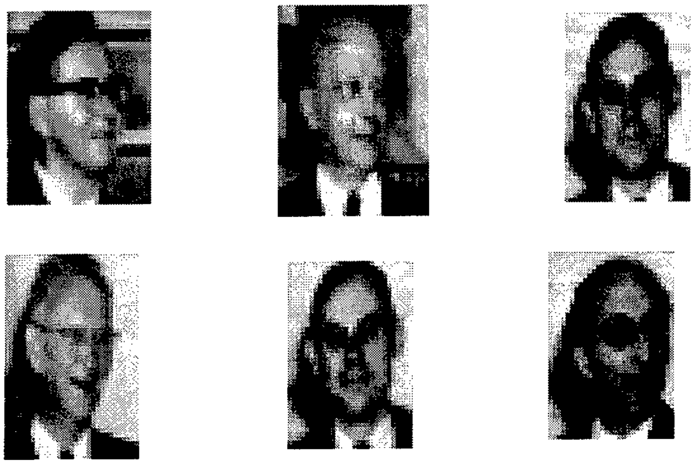


Figure 46. The six face objects memorized for Trial Set 4.

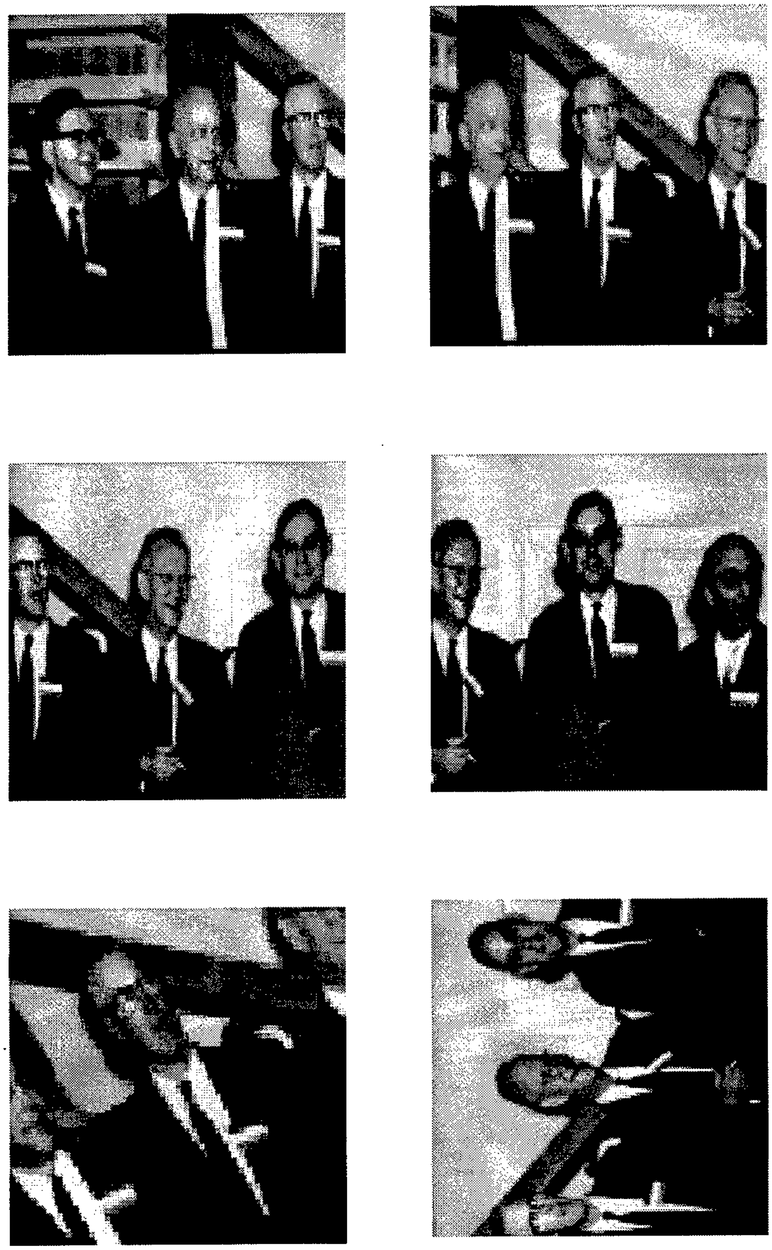


Figure 47. The original four target images tested against for Trial Set 4 (top two rows) and examples of two of the variants (bottom row).

A.5 Trial Set 5 Data

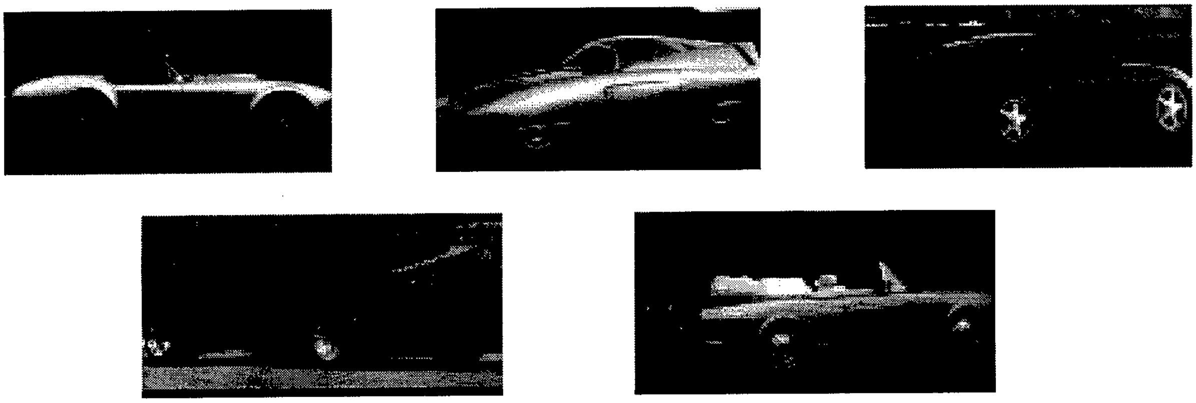


Figure 48. The five car objects memorized for Trial Set 5.

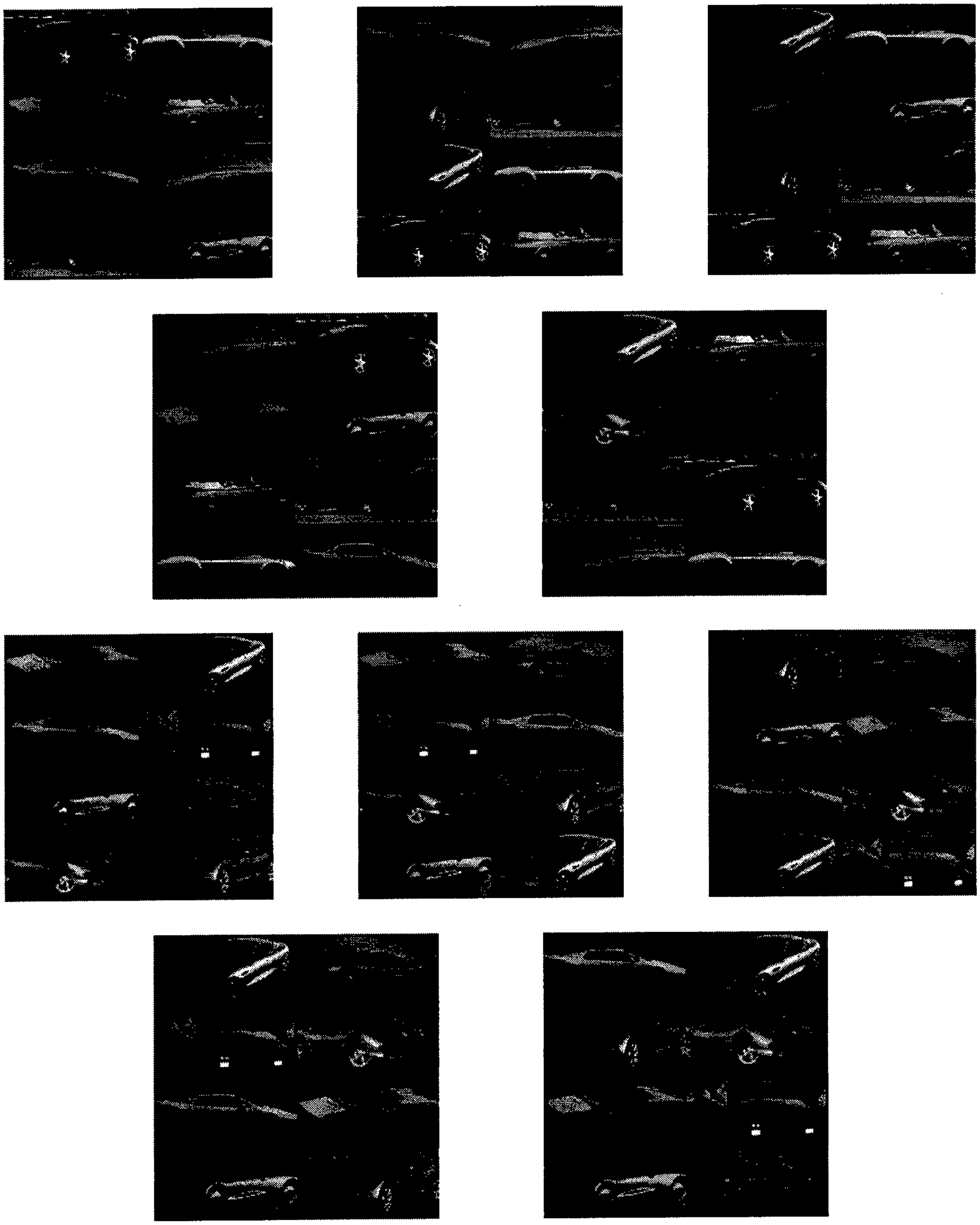


Figure 49. The original 10 target images tested against for Trial Set 5.

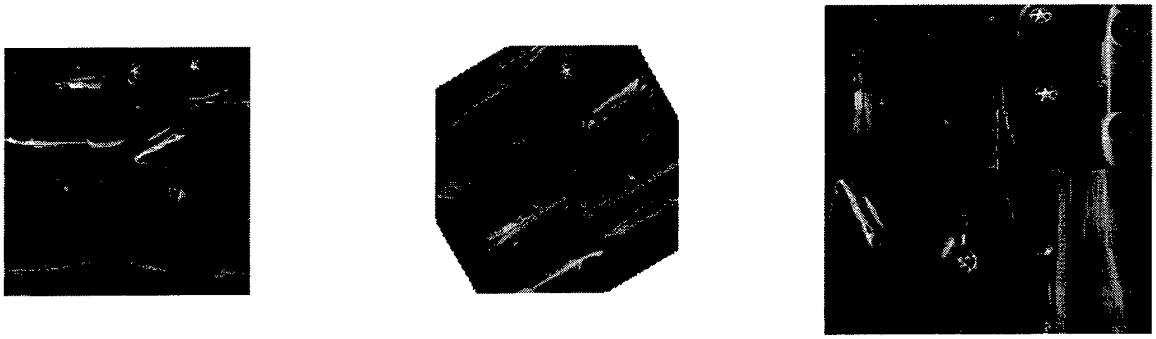
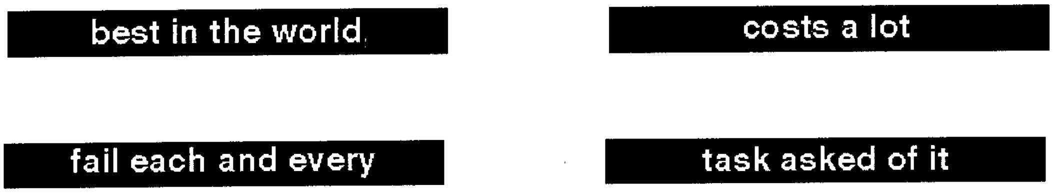


Figure 50. Three examples of the variants used for Trial Set 5 testing.

A.6 Trial Set 6 Data



best in the world

costs a lot

fail each and every

task asked of it

Figure 51. The four text objects memorized for Trial Set 6. The same car objects used for Trial set 5 were also memorized here.

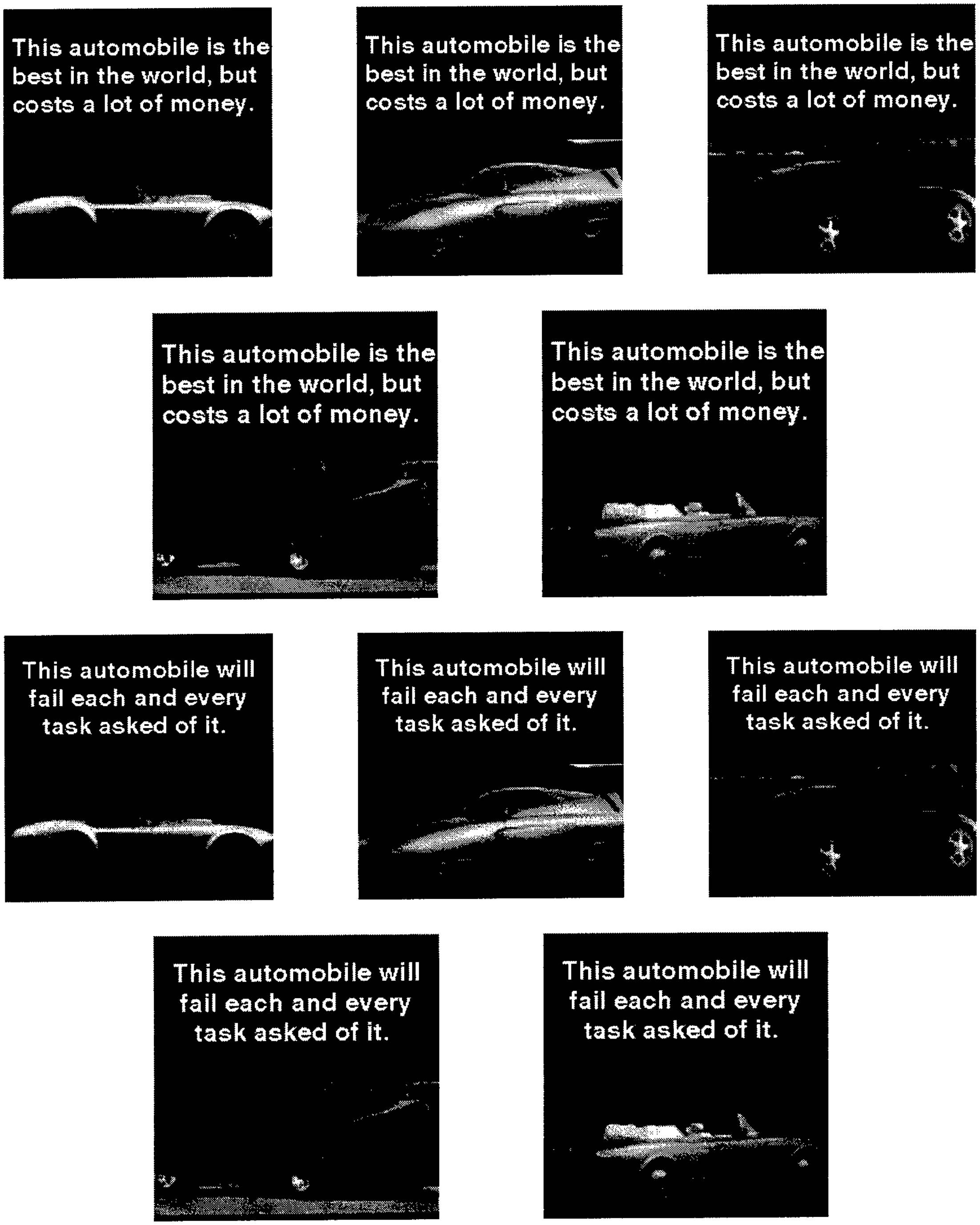

This automobile is the best in the world, but costs a lot of money.
This automobile is the best in the world, but costs a lot of money.
This automobile is the best in the world, but costs a lot of money.
This automobile is the best in the world, but costs a lot of money.
This automobile is the best in the world, but costs a lot of money.
This automobile will fail each and every task asked of it.
This automobile will fail each and every task asked of it.
This automobile will fail each and every task asked of it.
This automobile will fail each and every task asked of it.
This automobile will fail each and every task asked of it.

Figure 52. The original 10 target images tested against for Trial Set 6.

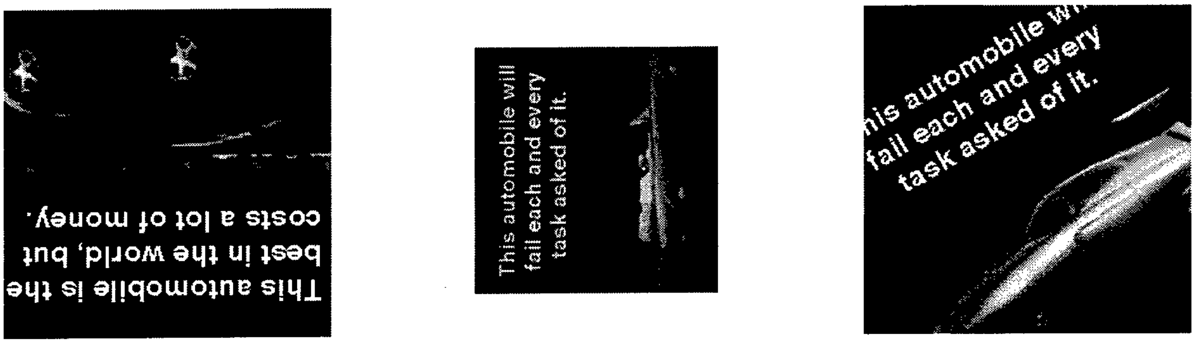


Figure 53. Three examples of the variants used for Trial Set 6 testing.

Appendix B. Video Mosaic Processing

B.1 Introduction

Living in the age of information has resulted in the generation and availability of vast stores of digital data. This data may be in the form of anything from simple textual data to complex video imagery, but regardless of the form, the problem of determining which data is relevant to a particular task is a daunting one. A prominent current example is found in the task of analyzing video imagery produced by Unmanned Aerial Vehicles (UAVs). To examine video footage for the presence of a particular target of interest, an automatic target detection algorithm may be applied to each frame individually, but may fail utterly because information necessary to determine the presence of the target was spread across multiple frames. Alternatively, a human observer may be used to examine the video playback, as we are quite effective at fusing information in the dynamic imagery internally, producing the illusion of continuity. However, the time and expense of dedicating man-hours to watching the playback quickly becomes prohibitive. Additionally, though we are quite good at performing such tasks, the observer suffers the human frailties of distraction and exhaustion as the length of the playback session increases.

A method is therefore desired to combine the advantages of automation with those of using a human observer. SCAN-IT is proposed as a method to perform this task. The fundamental concept is to convert a sequence of video frames into a single video mosaic that contains the 'important' information available in any of the individual frames. Because SCAN-IT is motivated by human physiology, the information in the mosaic is what would be considered important to the human visual perception system. The two primary tasks needing to be performed, then, are registration of the video frames to each other, and subsequent determination of the details in each frame that should be preserved in the mosaic.

The development and functional capabilities of SCAN-IT have been extensively detailed elsewhere in this document, so no more details will be offered here. The following sections will describe the use of SCAN-IT to perform the registration and fusion described above.

B.2 Frame Registration and Fusion

Using SCAN-IT for frame registration is actually a similar problem to using it for object recognition as reported on in Chapter V. In this case, the "object" being memorized will be the contents of a video frame, and recognition will result when a subsequent video frame is presented to the system and the same, or similar, content is found. This process can be thought of as the automated determination of fiducial landmarks in one frame, followed by the subsequent discovery of the same landmarks in the next frame. Once those markers are found, calculation of the translation and rotation required to register the frames is trivial.

An example of the data used to demonstrate the registration capability is provided in Figure B.2, which shows two different frames from video imagery captured by a UAV. a and b show the two frames prior to any processing, c shows the saccadic scan of the first frame during the memorization process, and d shows the saccadic scan during the search of the second frame. The dashed line in Figure B.2d shows the path as the vision model searched for a 'familiar' landmark, and the solid line shows the discovery and replication of the memorized path over the new frame.

Once the location of the landmarks (fixation points) from the first frame are found on the second frame, the rotation and translation for proper registration can be calculated and the two frames can be aligned. This process is repeated for all frames, and at its conclusion, the relative spatial location of all frames can be determined. However, ambiguity is introduced at every spatial location at which two or more frames overlap. Because simply overlaying each frame over the previous frame will result in the loss of any information available in previous frames, a method is desired to fuse the video data such that the "important" information in each frame is maintained. The physiological basis of SCAN-IT prompts one to define these important visual features as those producing a large response within the simulated visual field. The fusion is thus performed by weighting the contribution of each of the original video frames I on a pixel-by-pixel basis. The weighting for each spatial location is determined by

$$W_{x,y}^{I_i} = \frac{|E_{x,y}^{f_i}|}{\sum_{j=1}^{F} |E_{x,y}^{f_j}|} , \tag{17}$$

where $W_{x,y}^{I_i}$ is the weight to be applied to the magnitude at spatial location (x, y) of the unprocessed frame I_i, $|E_{x,y}^{f_i}|$ is the magnitude of the edge response at spatial location (x, y) of the registered edge map image f_i, and F is the number of frames being processed. Thus the contribution of each frame (at a particular (x, y) location) is directly proportional to the contribution of the corresponding edge map. Given input frames I_i then, the value at each location in the resultant output mosaic O is given by

$$O_{x,y} = \sum_{i=1}^{F} [W_{x,y}^{I_i} \cdot I_{i,x,y}] \tag{18}$$

Figure 55 shows an example of using the process described above to register and fuse six frames of the UAV data. Though this is obviously a simple demonstration, it is considered illustrative of the potential of the technique. It is not currently reasonable to implement the process in a real time manner (interpreted Matlab code on a Sun UltraSparc 1 requires approximately 1 second per saccade), but work is ongoing to increase efficiency. Rotational adjustments are also somewhat limited by the number of oriented Gabor filters used in SCAN-IT, but as efficiency increases, the number of orientations (filters) can be increased.

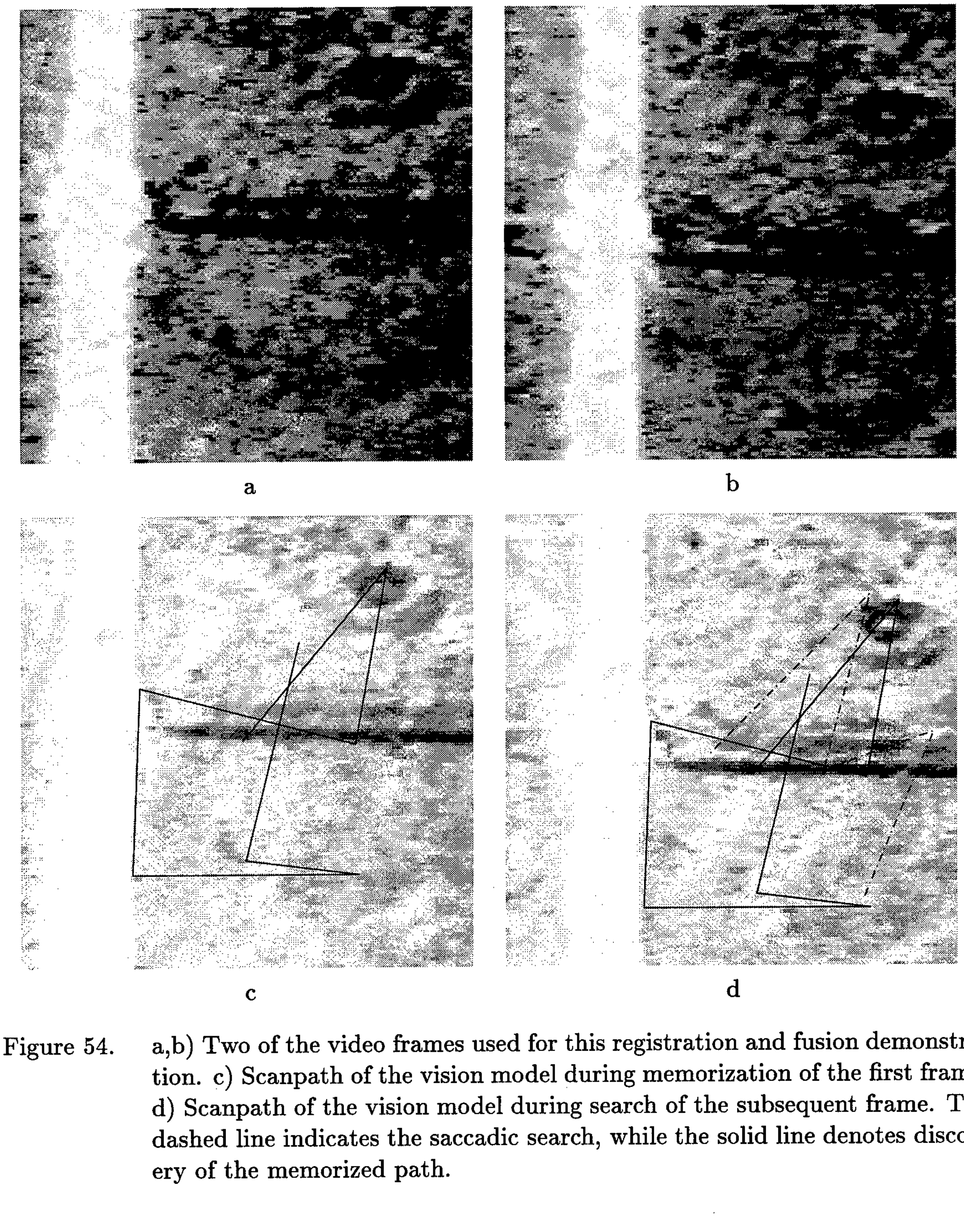


Figure 54. a,b) Two of the video frames used for this registration and fusion demonstration. c) Scanpath of the vision model during memorization of the first frame. d) Scanpath of the vision model during search of the subsequent frame. The dashed line indicates the saccadic search, while the solid line denotes discovery of the memorized path.

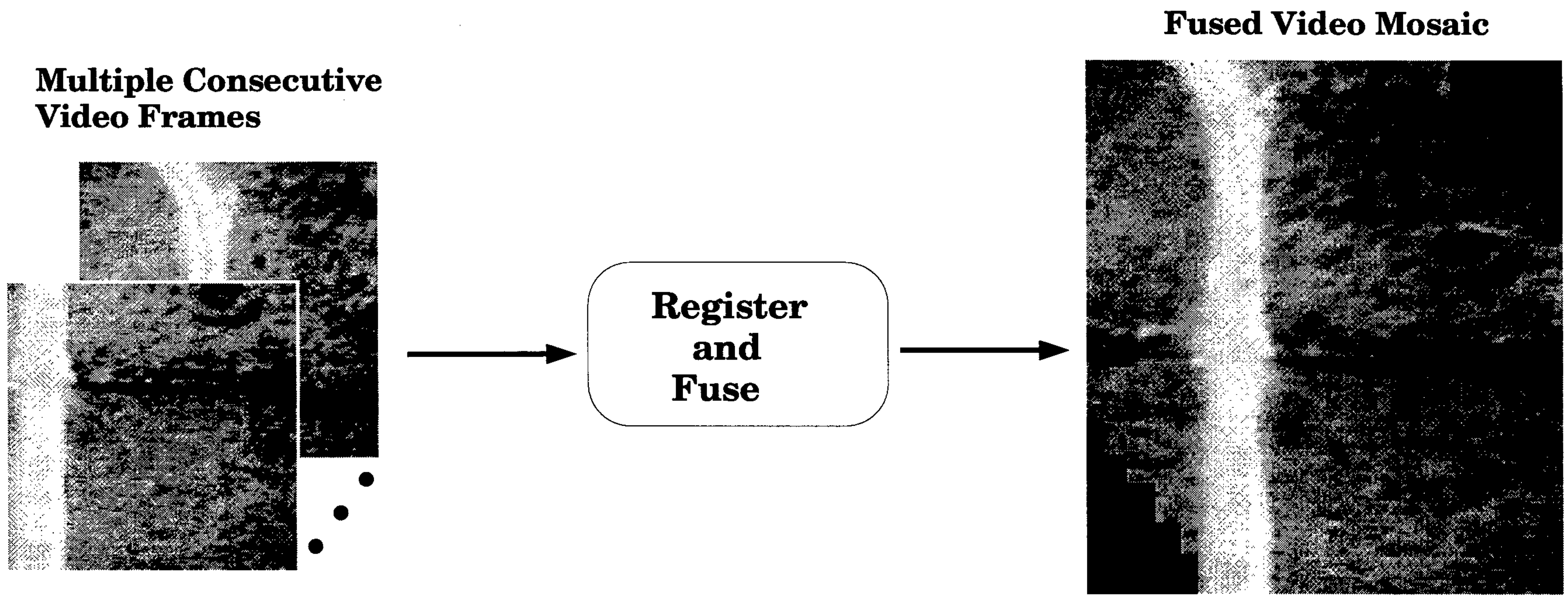


Figure 55. Registration and fusion of six frames of video data.

B.3 Conclusion

For the work presented here, a method of reducing multiple frames of video imagery into a single video mosaic has been introduced. The frames of the original imagery are first registered to each other using a model based on characteristics of the human visual system to 'memorize' each frame and then 'recognize' where the next frame properly aligns with it. Once this registration process is complete, the mosaic is produced by fusing the important (to the human visual system) information from each of the individual frames into a single, coherent image. Such a data reduction method will greatly reduce the manpower needed for analysis by eliminating the need to actively observe either dynamic or frame-by-frame playback of the imagery. At the same time, important information that may have only been present in a very few frames will be retained.

Bibliography

1. Ardizzone, Edoardo, et al. "Content Based Indexing of Image and Video Databases by Global and Shape Features." *International Conference on Pattern Recognition.* August 1996.
2. Ballard, D. H., et al. "Deictic codes for the embodiment of cognition," *Behavioural and Brain Sciences* (1997).
3. Bell, Anthony J. and Terrence J. Sejnowski. "Edges are the Independent Components of Natural Scenes." *Advances in Neural Information Processing Systems 9*, edited by Michael C. Mozer, et al. May 1997.
4. Boynton, Robert M. *Human Color Vision.* New York, NY: Holt, Rinehart and Winston, 1979.
5. Burt, P. J. "Smart Sensing Within a Pyramid Vision Machine." *Proceedings of the IEEE 76.* 1006–1015. 1988.
6. Carrasco, J. and K. S. Frieder. "Cortical Magnification Neutralizes the Eccentricity Effect in Visual Search," *Vision Research*, *37(1)*:63–82 (1997).
7. Cascia, Marco La and Edoardo Ardizzone. "JACOB: Just A Content-Based Query System for Video Databases." *Proceedings of the ICASSP.* May 1996.
8. Cavanagh, Patrick. "Functional Size Invariance is not Provided by the Cortical Magnification Factor," *Vision Research*, *22*:1409–12 (1982).
9. Das, Sanjoy, et al. "A Distributed Model of the Saccadic System: The Effects of Internal Noise," *Neurocomputing*, *11*:245–69 (1996).
10. Daugman, J. G. "Two Dimensional Spectral Analysis of Cortical Receptive Field Profiles," *Vision Research*, *20*:847–56 (1980).
11. Daugman, J. G. "Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36(7)*:1169–79 (1988).
12. Daugman, John G. "Six Formal Properties of Two-Dimensional Anisotropic Visual Filters: Structural Principles and Frequency/Orientation Selectivity," *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(5):882–7 (1983).
13. Daugman, John G. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America*, *2*(7):1160–1169 (1985).
14. Didday, R. L. and M. A. Arbib. "Eye Movements and Visual Perception: A Two Visual System Model," *International Journal of Man-Machine Studies*, *7* (1975).
15. Duda, Richard O. and Peter E. Hart. *Pattern Classification and Scene Analysis.* New York, NY: John Wiley and Sons, 1973.

16. Egly, Robert, et al. "Shifting Visual Attention Between Objects and Locations: Evidence from Normal and Parietal Lesion Subjects," *Journal of Experimental Psychology: General*, *123:2* (1994).
17. Findlay, John M. "Global Visual Processing for Saccadic Eye Movements," *Vision Research*, *22* (1982).
18. Gabor, D. "Theory of communication," *J IEE (London)*, *93*:429–457 (1946).
19. Gancarz, Gregory and Stephen Grossberg. "A Neural Model of the Saccade Generator in the Reticular Formation," *Neural Networks*, *11* (1998).
20. Gaskill, Jack D. *Linear Systems, Fourier Transforms, and Optics*. New York: John Wiley & Sons, 1978.
21. Grossberg, S., et al. "A Neural Network Architecture for Preattentive Vision," *IEEE Transactions on Biomedical Engineering*, *36* (1989).
22. Gupta, Amarnath. *Visual Information Retrieval Technology; A Virage Perspective*. Technical Report, 177 Bovet Road; Suite 520;San Mateo, CA 94402: Virage, Inc, 1997.
23. Hafner, James, et al. "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17, No. 7* (July 1995).
24. Hall, Charles F. *Digital Color Image Compression in a Perceptual Space*. PhD dissertation, University of Southern California, 1978.
25. Harman, D., editor. *The First Text Retrieval Conference*. Washington DC: NIST, 1992.
26. Harman, D., editor. *Proceedings of the Second Text Retrieval Conference*. Washington DC: NIST, 1994.
27. He, Peiyuan and Eileen Kowler. "The Role of Location Probability in the Programming of Saccades: Implications for 'Center-of-Gravity' Tendencies," *VIsion Research*, *29, No. 9* (1989).
28. Hecht-Nielsen, R. and Y. T. Zhou. "VARTAC: a Foveal Active Vision ATR System," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17, No. 7* (July 1995).
29. Howarth, R. J. and H. Buxton. *Selective Attention in Dynamic Vision*. Technical Report 610, Queen Mary and Westfield College, 1992.
30. Hubel, D. H. and T. N. Wiesel. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *Journal of Physiology*, *160* (1962).
31. Hubel, D. H. and T. N. Wiesel. "Receptive Fields and Functional Architecture of Monkey Striate Cortex," *Journal of Physiology*, *195* (1968).
32. Hubel, D. H. and T. N. Wiesel. "Sequence, Regularity, and Geometry of Orientation Columns in the Monkey Striate Cortex," *Journal of Comparative Neurology*, *158* (1974).

33. Hubel, David H. *Eye, Brain, and Vision.* New York, New York: Scientific American Library, 1995.

34. Indiveri, Giacomo, et al. "A Recurrent Neural Architecture Mimicking Cortical Preattentive Vision Systems," *Neurocomputing*, *11* (1996).

35. Jain, Ramesh, et al. "Similarity Measures for Image Databases," *SPIE*, *2420* (February 1995).

36. Jones, G. J. F., et al. "Video Mail Retrieval: The Effect of Word Spotting Accuracy on Precision." *Inernational Conference on Acoustics, Speech and Signal Processing1.* 1995.

37. Jones, J. P. and L. A. Palmer. "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophys.*, *58*:1233–1258 (1987).

38. Kelly, Patrick M, et al. "Efficiency Issues Related to Probablility Density Function Comparison," *SPIE*, *2670* (1996).

39. Kelly, Patrick M, et al. "Query by Image Example: the CANDID Approach," *SPIE*, *2420* (1995).

40. Klein, Raymond and Mark Farrell. "Search Performance without Eye Movements," *Perception and Psychophysics*, *46:5* (1989).

41. Kosslyn, S. M., et al. "Components of high-level vision: A cognitive neuroscience analysis and account of neurological syndromes," *Cognition*, *34* (1990).

42. Kowler, E., et al. "The Role of Attention in the Programming of Saccades," *Vision Research*, *35* (1995).

43. Kowler, E. and S. Anton. "Reading Twisted Text:Implications for the Role of Saccades," *Vision Research*, *27* (1987).

44. Levi, Dennis M., et al. "Vernier Acuity, Crowding and Cortical Magnification," *Vision Research*, *25*(7) (1985).

45. Liu, Fang and Rosalind W. Picard. "Periodicity, Directionality, and Randomness: Wold Features for Image Modieling and Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18, No. 7* (July 1996).

46. Manjunath, B S and W Y Ma. "Texture Features for Browsing and Retrieval of Image Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (November 1996).

47. Manmatha, R., et al. "Word Spotting: A New Approach to Indexing Handwriting." *Proceedings of the IEEE Computer VIsion and Pattern Recognition COnference (CVPR).* June 1996.

48. Mannos, James L. and David J. Sakrison. "The Effects of a Visual Fidelity Criterion on the Encoding of Images," *IEEE Transactions on Information Theory*, *IT-20*(4) (July 1974).

49. Mao, Jianchang and Anil K. Jain. "Texture Classification and Segmentation using Multresolution Simultaneous Autoregressive Models," *Pattern Recognition*, *25, no. 2* (1992).

50. Martin, Curtis E. *Perceptual Fidelity for Digital Color Imagery*. PhD dissertation, AFIT, 1996.

51. Mehrotra, R. and J. Gary. "Feature-Index-Based Similar Shape Retrieval." *Visual Database 3: Proc. of the 3rd IFIP 2.6 Working Conf. on Visual DB Systems* edited by S. Spaccapietra and R. Jain, Chapman and Hall, 1995.

52. Mehrotra, R. and J. E. Gary. *A Technique for Retrieval of Similar Images of Three-Dimensional Objects*. Technical Report, University of Missouri-St Louis; Mathematics and Computer Science Department, 1996.

53. Mishkin, M., et al. "Object Vision and Spatial Vision: Two Cortical Pathways," *Trends in Neurosciences*, *6* (1983).

54. Niblack, W, et al. "The QBIC Project: Querying Images by Content using Color, Texture, and Shape," *SPIE*, *1908* (1993).

55. Noton, D. and L. Stark. "Scanpaths in Eye Movements During Pattern Recognition," *Science*, *171* (1971).

56. Ogle, Virginia E. and Michael Stonebraker. "Chabot: Retrieval from a Relational Database of Images," *IEEE Computer* (1996).

57. Olshausen, Bruno A. and David J. Field. "Wavelet-like Receptive Fields Emerge from a Network that Learns Sparse Codes for Natural Images," *Nature*, *381* (1996).

58. Parsons, Thomas W. *Voice and Speech Processing*. New York, NY: McGraw-Hill Inc, 1987.

59. Rao, Rajesh P. N. and Dana H. Ballard. "An Active Vision Architecture based on Iconic Representations," *Artificial Intelligence (Special Issue on Vision)*, *78*:461–505 (1995).

60. Rao, Rajesh P. N. and Dana H. Ballard. "Learning Saccadic Eye Movements Using Multiscale Spatial Filters," *Advances in Neural Information Processing System 7*, 893–900 (1995).

61. Rao, Rajesh P. N., et al. *Eye Movements in Visual Cognition*. Technical Report 97.1, National Resource Laboratory for the Study of Brain and Behavior, University of Rochester, March 1997.

62. Rimey, Raymond D. and Christopher M. Brown. "Controlling Eye Movements with Hidden Markov Models," *International Journal of Computer Vision*, *7:1*:47–65 (1991).

63. Rimey, Raymond D. and Christopher M. Brown. "Task-Specific Utility in a General Bayes Net Vision System," *Computer Vision and Pattern Recognition (CVPR)*, *92*:142–147 (1992).

64. Rodieck, R. W. "Quantitative Analysis of Cat Retinal Ganglion Cell Response to Visual Stimuli," *Vision Research*, *5*:583–601 (1965).

65. Rogers, Steven K. and Mathew Kabrisky. *An Introduction to Biological and Artificial Neural Networks for Pattern Recognition*. SPIE Optical Engineering Press, 1991.

66. Rybak, I. A., et al. *A Model of Attention-Guided Visual Perception and Recognition*. Technical Report, Wilmington, DE 19880-0328: E. I. du Pont de Nemours and Co., Central Research Department, March 1997.

67. Rybak, Ilya A., et al. "Behavioral Model of Visual Perception and Recognition," *SPIE*, *1913* (1993).

68. Saarinen, Jukka. "Shifts of Visual Attention at Fixation and Away from Fixation," *Vision Research*, *33:8* (1993).

69. Sclaroff, S. and A. Pentland. "Modal Matching for Correspondence and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17, No. 6* (June 1995).

70. Sekuler, R. and R. Blake. *Perception*. New York, NY: McGraw-Hill, Inc, 1994.

71. Swain, Michael J. and Dana H. Ballard. "Color Indexing," *International Journal of Computer Vision*, *7:1* (1991).

72. Swokowski, Earl W. *Elementary Functions with Coordinate Geometry*. Boston, MA: Prindle, Weber, and Schmidt, Inc, 1971.

73. Tou, J. T. and R. C. Gonzalez. *Pattern Recognition Principles*. Reading, MA: Addison-Wesley PUblishing Company, 1974.

74. Turner, M.R. "Texture discrimination by Gabor functions," *Biological Cybernetics*, *55*:71–82 (1986).

75. Van Essen, D. "Functional Organization of Primate Visual Cortex." *Cerebral Cortex, Vol. 3* edited by A. Peters and E. G. Jones, New York, NY: Plenum Press, 1985.

76. Wahl, F., et al. "Block Segmentation and Text Extraction in Mixed Text/Image Documents," *Computer Vision Graphics and Image Processing*, *20* (1989).

77. Wandell, Brian A. *Foundations of Vision*. Sunderland, MA: Sinauer Associates, Inc, 1995.

78. Wässle, H., et al. "Functional Architecture of the Mammalian Retina," *Vision Research*, *30* (1990).

79. Wilson, Terry A., et al. "Perceptual-based hyperspectral image fusion using multiresolution analysis," *Optical Engineering*, *34*(11) (1995).

80. Wu, Victor, et al. "Finding Text in Images," *2nd ACM International Conference on Digital Libraries* (1997).

81. Yarbus, A. L. *Eye Movement and Vision*. New York, NY: Plenum Press, 1967.

82. Zeki, S. *A Vision of the Brain*. Blackwell Scientific Publications, 1993.

83. Zeki, S. "The visual image in mind and brain." *Mind and Brain* 27–39, W.H. Freeman and Company, 1993.

Vita

Captain John Keller enlisted in the US Air Force and entered active duty in 1981. During the next 6 years he served as a Ground Radio Communication Equipment Repairman, rising in rank from Airman Basic to Staff Sergeant. During this time, he was granted an Associate's of Science degree from the Community College of the Air Force and an Associate's degree in Management from City Colleges of Chicago.

In 1987, Captain Keller was selected to attend the University of Florida as a full-time student under the auspices of the Airman Education and Commissioning Program. He was graduated from that institution in 1989 with a Bachelor of Science degree in Electrical Engineering, immediately after which he departed for Lackland Air Force Base to attend Officer Training School. After being commissioned as a 2nd Lieutenant, he worked for two and a half years as a project engineer/manager for the Airborne Warning and Control System (AWACS) program at Tinker Air Force Base, Oklahoma. During this period, he was awarded a Master's degree in Computer Resource Management from Webster University.

His next military assignment took him to Wright-Patterson Air Force Base to attend the Air Force Institute of Technology (AFIT), from which he was graduated with a Master of Science degree in Electrical Engineering in 1993. This educational stint was followed by a Test and Evaluation role in Melbourne, Florida, with Captain Keller serving the next two and a half years as a Flight Test Engineer on the aircraft component of the Joint Surveillance Target Attack Radar System (Joint STARS). His next assignment brought him back to AFIT in 1996 to pursue a doctoral program. The follow-on assignment from this AFIT program is with the Air Force Research Laboratory in Rome, New York.

Permanent address: 3423 Calumet Dr
Orlando, FL 32810

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
June 1999

3. REPORT TYPE AND DATES COVERED
PhD Dissertation

4. TITLE AND SUBTITLE
SCAN-IT: A Computer Vision Model Motivated by Human Physiology and Behavior

5. FUNDING NUMBERS

6. AUTHOR(S)
John G. Keller

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/DS/ENG/99-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Major Steve Matechik
AFRL/IFEC
32 Brooks Road
Rome, NY 13441
(315)330-4426

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This dissertation details the development of a new computational vision model motivated by physiological and behavioral aspects of the human visual system. Using this model, intensity features within an artificial visual field of view are extracted and transformed into a simulated cortical representation, and a saccadic guidance system scans this field of view over an object within an image to 'memorize' that object. The object representation is thus stored as a sequence of feature matrices describing sub-regions of the object. A new image can then be searched for the object (possibly scaled and rotated), where evidence of its presence is accumulated by finding sub-regions in the new image similar to those stored and with the same relative spatial configuration. A set of over 450 experimental trials demonstrates the model is capable of memorizing and then recognizing arbitrary objects within arbitrary images, as well as correctly rejecting images that do not contain the memorized object. A new context-based recognition paradigm is introduced that solves the problem of a priori assignation of recognition threshholds, and also can be generalized to solve threshholding problems commonly found in pattern recognition environments. A demonstration is provided of the model's applicability to real-world problems by memorizing a face and text string, and then successfully searching a video sequence for their presence.

14. SUBJECT TERMS
Human Visual System, Computer Vision, Gabor, Saccadic Behavior, Active Vision

15. NUMBER OF PAGES
114

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Standard Form 298 (Rev. 2-89) (EG)
Prescribed by ANSI Std. 239.18
Designed using Perform Pro, WHS/DIOR, Oct 94