A MODEL OF ACTIVE AND ATTENTIVE VISION

by

Çağatay Soyer

B.S. in Electrical and Electronic Engineering, Boğaziçi University, 1993
M.S. in Biomedical Engineering, Boğaziçi University, 1995

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Doctor
of
Philosophy

Boğaziçi University
January, 2002

A MODEL OF ACTIVE AND ATTENTIVE VISION

APPROVED BY:

Doç. Dr. H. Işıl Bozma .....
(Thesis Supervisor)

Prof. Dr. Yorgo Istefanopulos .....
(Thesis Co-supervisor)

Doç. Dr. H. Özcan Gülçür .....

Doç. Dr. Mehmed Özkan .....

Prof. Dr. Aytül Erçil .....

DATE OF APPROVAL: 28.01.2002

A MODEL OF ACTIVE AND ATTENTIVE VISION

ABSTRACT

Biological vision systems explore their environment by allocating their resources to interesting parts of a scene, using both physical and mental attention mechanisms. The result of this active and attentive vision behavior is a sequence of images obtained from different spatial locations at different times. However, temporal processes and integration mechanisms in the brain enable us to interpret this information and perceive a stable image of the environment. While models of such attention and perception mechanisms are invaluable to understand human vision, they are also increasingly used and improved by robotics and artificial intelligence researchers to achieve human-like performance. In a similar attempt, we propose a new and complete model of active vision behavior, based on confirmed biological evidence where available. The model consists of an attention system, temporal image sequence processing algorithms and an integrative visual memory. All components of the model are implemented on our mobile robot APES. Gaze control, sequence based scene recognition and visual integration tasks are assumed during experiments. Results of gaze control experiments clearly demonstrate a human-like selective attention behavior, which can be fully controlled by a number of parameters. In recognition and integration tasks, simple and complex scenes were successfully modeled and classified. Furthermore, our work on attentional image sequences raised a number of interesting questions, some of which have been answered in this thesis.

Keywords: Selective attention, robot vision, active vision, attentive vision, eye movements, temporal recognition, scene classification.

AKTİF VE İLGİYE DAYALI GÖRME MODELİ

ÖZET

Biyolojik görme sistemleri fiziksel ve zihinsel ilgi mekanizmaları kullanarak bilgi işlem kaynaklarını ilginç görüntü bölgelerine yoğunlaştırırlar. Bu aktif ve ilgiye dayalı görme davranışının sonucunda farklı alanlardan farklı zamanlarda alınmış bir dizi görüntü elde edilir. Ancak beyindeki zamansal işlem ve birleştirme mekanizmaları bu veriyi değerlendirmemizi ve ortamın sabit, durağan bir görüntüsünü algılamamızı sağlar. Bu tür ilgi ve algı mekanizmalarına ait modeller insan görme sisteminin anlaşılmasını sağlarken bir yandan da, insana yakın bir başarıma ulaşmak amacıyla, robot ve yapay zeka araştırmacıları tarafından kullanılmakta ve geliştirilmektedir. Benzer bir girişimle bu raporda, varolduğunda biyolojik kanıtlara dayanan, yeni ve tam bir aktif görme davranışı modeli öneriyoruz. Model bir ilgi sisteminden, zamansal görüntü dizilerini işleyen algoritmalardan ve birleştirici bir görsel bellekten oluşmaktadır. Modelin tüm bileşenleri hareketli robotumuz APES üzerinde uygulanmıştır. Deneyler sırasında bakış kontrolü, dizilere dayalı görüntü tanıma ve görsel birleştirme işleri yapılmıştır. Bakış kontrolü deneylerinde açıkça belli parametrelerle kontrol edilebilen insansı seçici ilgi davranışı gözlenmiştir. Tanıma ve birleştirme deneylerinde çeşitli basit ve karmaşık görüntüler başarıyla modellenmiş ve sınıflandırılmıştır. Ayrıca ilgiye dayalı görüntü dizileri üzerindeki çalışmalarımız bir kısmı bu raporda cevaplanan bazı ilginç soruları ortaya çıkarmıştır.

Anahtar Sözcükler: Seçici ilgi, robotla görme, aktif görme, ilgiye dayalı görme, göz hareketleri, temporal tanıma, görüntü tanıma.

ACKNOWLEDGEMENTS

My family and close friends have been putting up with my research activities for a long time. I would like to thank them for their patience and support.

I would like to thank my supervisor Associate Professor Işıl Bozma for her technical contribution and support, and my associate supervisor Professor Yorgo Istefanopulos for offering his experience in a variety of technical and administrative issues. I also thank the thesis committee for evaluating progress reports and attending my presentations, and our institute secretary Mrs. Berrin Kocayurt for her administrative support.

I would also like to thank TUBITAK, Bogazici University Research Fund, Boğaziçi University Foundation, ICAR'97 and IROS 2000 organization committees for their financial support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

I_v^t	Visual field image at time t.
I_f^t	Fovea image at time t.
I_f^c	Candidate fovea image.
v	Fovea feature value.
o^t	Observation about the state of the fovea.
Ω_m	Set of values of m^{th} visual primitive.
I_h^t	Inhibition region image at time t.
f_m	Feature value operator.
a	Attention function.
O^T	Observation sequence.
A^l	State transition probability matrix for l^{th} scene.
A_l	l^{th} proposition about a scene.
l^*	Correct classification of a scene.
L	Set of possible values for a scene classification.
C_d	Set of previously fixated foveas, FIFO memory.
T_l	Feature transition frequency matrix for scene l.
ω	Weight of evidence function.
m	Basic probability number.
s_l	Simple support function focused on A_l.
s_l^i	Instantaneous support for A_l.
s_l^t	Temporal support for A_l.
β	Set of bubble points.
B	Bubble matrix.
ρ	Bubble function.

1. INTRODUCTION

Understanding human vision is a joint effort by biological and computational vision researchers. At the intersection of these two practically distinct but theoretically similar goals, vision implants, prostheses, humanoid robots, simulation models, new clinical equipment and treatment techniques are capitalizing on the same research. This thesis is an effort at this intersection, towards modeling attention, memory, scene representation and recognition mechanisms in biological vision and implementing them on an attentive robot vision system.

Mathematical and conceptual models contribute to our understanding of biological systems. Furthermore they are also widely employed in robotics and automation to mimic biological performance in artificial systems. In this respect, artificial vision research faces a big challenge, as the vision system is one of the most complex and poorly understood biological systems. Still, many structural and functional properties of primate vision have been successfully modeled and have lead to important developments in robot vision. Some of these are worth noting here to provide motivation for biologically inspired robotic systems.

The idea of representing visual scenes using edges was a direct consequence of the discovery of neurons responding to edges in the visual cortex [1,2,3,4,6,7,8,9]. Edge detection later became a starting point for many computer vision algorithms [10,11,12,13]. Hubel and Wiesel's exciting discovery of cortical cells responding to edges in different orientations [4] and recent experiments by Gallant et al. [2,3] inspired the use of multi-orientation, multi-frequency feature detectors instead of classical horizontal and vertical operators. Topographical organization and specialization in the visual cortex was discovered by Zeki and a number of other physiologists [13]. These discoveries confirmed that different modalities of the scene were processed separately and contributed to scene understanding through complex reentrant connections in the cortex. Other concepts and tools like neural networks, learning algorithms, filtering and transformation techniques,

emergent behavior, were all either inspired or confirmed by biological findings [10,12, 13,15,16,17,18].

Later in early 80s artificial vision research concentrated on another important discovery by psychologists and physiologists. While psychological experiments as early as the 1970s suggested that vision was a serial process [18,19,20,21], physiologists verified that there was a high concentration of photoreceptor cells on the central fovea region of the retina - which had the consequence of necessitating eye motions in order to analyze a scene in detail [7,22]. This enabled to avoid detailed processing of non-relevant visual data. Further research has revealed both overt attention mechanisms characterized by head and eye motions and unconscious covert attention mechanisms [6,13,18,19,23,24,25,27,29, 31,34].

The discovery of attention effects and serial processing was closely related to major problems of artificial vision including real time operation and performing complex visual tasks in a dynamic environment. A new vision paradigm made use of these findings and was called *active vision* or *animate vision* [35,36,37,38,40,41,42,43,44]. In this thesis a third term, *attentive vision*, is also used to emphasize the fact that an attention mechanism is the major difference between classical and active vision systems.

1.1 Problem Statement

Working with active vision systems first requires the solution of selective attention problem by simulating the structure and movements of the eye. A retina model providing fovea-periphery distinction and an electromechanical system, which can be controlled to direct attention, must be developed. Pre-attentive and attentive features must be defined and extracted according to the given task. Intelligent algorithms for selecting the next fixation point are also required to maximize the information content of the generated attentional sequence.

In order to be able to use active vision systems in real world vision tasks, the output of the system – the attentional image sequence – must be processed appropriately. As a basic requirement the system must be able to learn, generalize, and mathematically represent the stream of images of different spatial locations fixated using the selective attention mechanism. Then, these models can be used to recognize sequences or parts of sequences, which will in turn enable the system to make decisions based on visual observation. The sequence modeling and recognition algorithms should also have temporal properties, such that a recognition task can be completed at any time while looking at a scene, if sufficient information is collected.

Another interesting problem in active vision is environment modeling during selective attention, usually by a system which is both active and mobile. This problem also has common properties with the problem of visual integration over fixations in cognitive psychology. The basic question is how to combine information collected during different fixations, such that a stable image or model of the environment is always available for higher level functions. Humans perform this integration unconsciously and perceive a stable image of their environment in spite of continuous eye movements, but the underlying mechanisms are still unknown.

1.2 Related Literature

1.2.1 Biology of Vision and Attention

1.2.1.1 Processing in the Retina

Since the 1950s many researchers have worked on the structure and functions of the retina, making it the best understood component of the mammalian visual system [3,4,7, 8,9,22]. Unlike other receptors in the body, the retina is not a peripheral organ, but a part of the central nervous system which has direct connections to brain structures. Light sensitive receptors on the retina - cones and rods for day and night vision respectively - are the first neural elements in the visual pathway. Visual information from the receptors is processed

and projected to higher centres of vision by bipolar, horizontal, amacrine, and ganglion cells, found in three layers of the retina. The final projection neurons are the ganglion cells which are organized in three parallel systems, X, Y, and W participating in high acuity vision, fast and crude analysis of the scene, and head and eye movements respectively. Ganglions convey the visual information, processed by the massive neural network of the retina, to the higher centers of vision like the lateral geniculate nucleus, superior colliculus, the primary visual cortex, and other brain structures. Bipolar cells are used to connect receptors to ganglions, and finally the horizontal and amacrine cells modulate the information flow from the receptor to bipolar to ganglion cells. Though processing performed by the retina is limited to blob detection in an on-centre off-surround or off-centre on-surround manner, an enormous amount of variation is built into this simple operation in terms of acuity, sensitivity, spatial integration, and processing speed. Moreover, the distribution of receptor cells on the retina is not uniform, but rather like a gaussian with a small variance, resulting in a loss of resolution as we move away from the optical axis of the eye. The small region of highest acuity around the optical axis is called the fovea, and the rest of the retina is called periphery. In very general terms, this fovea-periphery distinction in the retina brings about the need for fixating on regions of interest for detailed processing and thus, removes a great deal of redundancy.

1.2.1.2 Processing in the Visual Cortex

After developing a detailed map of the complex structure of the retina, research has been directed to the visual cortex and the relatively simple lateral geniculate nucleus acting as a distribution point between the retinal ganglions and the cortex. D.H.Hubel and colleagues accidentally found out that a large number of cells in the primary or striate cortex (also called area V1) - which they later named 'complex' - responded only to moving edges of a particular orientation [4,7,45]. Other cells called 'simple' were responding to still edges of a particular orientation placed properly in their receptive fields. Later, they also discovered cells with radially symmetric receptive fields responding to edges regardless of orientation, and cells with 'end-stopping' property for corner and curvature detection. When the receptive fields of these cells were considered, a topographical organization of connections, from the retina to lateral geniculates and from there to the primary and other

parts of the visual cortex, was observed. This means that the retinal image was represented in all of these centers, and the number of cells dedicated to representing some part of the retina were proportional to the number of receptors in that region. In this case the small fovea was represented by more cells in the cortex than the whole periphery. Another important discovery about the image representations in higher centers was related to the level of abstraction. As we move from the retina to the primary and other cortical regions, the features that the cells respond became more complex, and the receptive fields of these cells on the retina became larger and larger. For example in the primary visual cortex simple cells responded to lines of a particular orientation, more common complex cells responded to their movement, and some cells both simple and complex responded to specified corners and curvatures. In other cortical regions like V2, V3, V4, and V5 collectively called visual association cortex or prestriate cortex, different cells responded to color, motion, and orientation [4,13,45,47]. This specialization and abstraction of information also found evidence in clinical studies. However, contrary to many expectations, the cortical regions were not connected to a higher centre where the visual scene was perceived, rather they are found to be connected to each other by reentrant connections, as a result of recent work by S.Zeki and collegues [13].

1.2.1.3 The Oculomotor System

Humans explore the visual environment by directing their eyes to interesting stimuli. Due to the properties of human retina each time we look in a different direction we can only see the small foveal region in detail. The rest of the picture is not detailed enough to recognize any object or feature. The rapid motion (up to 400-600 deg/sec) between these *fixations* is called a *saccade*. The low speed of neural elements makes perception impossible during a saccade. Therefore we are practically blind during a saccade. As a result, the information obtained by the eye is a collection of images or snapshots of different spatial locations taken at different times. How this information is converted into a smooth, continuous, stationary image is currently one of the most interesting questions of vision research.

In many older studies usually five or six different types of eye movements controlled by the oculomotor system are listed. These are saccades, smooth pursuit movement, high frequency tremors, the opto-kinetic reflex, and vergence. However recent studies generally consider only three types of movements using different mechanisms. These are smooth pursuit movement, eye movements during maintained stable gaze, and saccades [18,19].

Maintained stable gaze: Maintained stable gaze is defined as the condition when we try to look at a stationary target. The behavior of the eyeball in this state is important to determine the exact retinal image seen by the higher vision centers in the brain. Although the eyes may look completely stationary during fixation, accurate measurements prove that in fact they make very small amplitude oscillations of varying frequencies.

In many studies high frequencey tremors (30-80 Hz) of small amplitudes (about 15 sec. arc) were observed to be superimposed on lower frequency (2-5 Hz) but higher amplitude (5-10 min. arc) oscillations. Also at intervals ranging from 0.2 to several seconds, small amplitude (5-10 min. arc) saccades (microsaccades) were observed. The result of maintained fixation studies was the discovery of the role of retinal image motion in vision. Any image artificially stabilized on the retina was found to fade away within a few seconds. In 1980s, almost 30 years after this observation, cortical cells responding only to moving stimuli were discovered in the visual cortex. It was also found that the stability of fixation and the mean position of the eyeball were not affected by the color of the target, the luminance of the target, the size of the target (priovided that it was small enough to fit in the fovea). It is now widely accepted that eye movements during maintained fixation are largely related to the operation of motion sensitive neurons. It is also thought that this continuous motion may be effective in reducing the latency of saccadic response.

Smooth pursuit movement: Unlike saccades smooth pursuit eye movements cannot be initiated voluntarily without a moving stimulus and they cannot be completely suppressed in the presence of a moving target. A well known pursuit or tracking movement called opto-kinetic nystagmus (OKN) has been widely used to detect anomalies of the

frontal lobe. OKN was previously thought to be a different reflex but recent studies showed that the mechanisms of smooth pursuit and OKN were essentially the same.

The behavior of the smooth pursuit system in the presence of more than one moving stimuli was one of the first questions about smooth pursuit. The straightforward answer was that the effects of all stimuli would be combined in some way but this was not true. Instead a selective attention mechanism determined the input to the oculomotor system. For example in the case of two moving targets one of the stimuli was chosen by selective attention and after a fixation saccade the target was smoothly tracked.

Another interesting property of smooth pursuit was its performance. In early measurements human eye seemed to respond very quickly to stimulus motion and tracking performance was also very good. As early as 1950s it was understood that human eye was using the observed error to predict target behavior and improve tracking performance. While researchers were trying to model this tracking performance by predicting controllers, tracking experiments using better measurement equipment proved that in fact the eye was moving almost as much as 300 ms. before the stimulus. Motion before the target clearly implied that smooth pursuit was also under control of cognitive factors like expectations and memory. In the case of random target motion the eye could also make a wrong start and then correct its motion based on the error. This anticipatory behaviour was first realized by Dodge in 1930s and it was better defined by Westheimer, Kowler, Steinman, Becker and other researchers in 80s and 90s.

As a result of anticipation, regular or periodic visual stimuli are tracked more accurately than stimuli moving in irregular patterns. The human vision system is able to predict target motion and guide eye movements accordingly. This is particularly important in analyzing and understanding human vision as any observed oculomotor response - either saccadic or smooth pursuit - is a combination of stimulus driven and anticipatory behaviors.

Saccades: Rapid eye motions between successive fixations are called saccades. Saccades are made 2-3 times every second and they are almost always initiated voluntarily. They are used to direct the fovea to an interesting location in the periphery

[18,19,28,29,31]. The accuracy of saccades decreases by the distance of the target to the current fixation. Human saccades are almost always larger than 0.5 degrees. In 1961 Rashbass proposed a hard wired saccadic *dead zone* based on this information, however it was found in 1973 that humans could in fact make smaller saccades when asked to do so and the dead zone was due to a reluctance to do so [18]. In other words the dead zone was a result of higher level cognitive processes.

For years researchers could not decide whether the saccade was a reflexive response to a stimulus or a voluntary behaviour. Studies in 1980s proved that saccadic responses were faster when the position of the stimulus could be guessed by the subject showing that saccades were programmed before the appearance of the stimulus. The flight path was also not changed after the saccade was initiated even if the target location changed. Another interesting result was that depending on the complexity and timing of the task, humans could choose to make faster but less accurate saccades and vice versa. All these studies showed that saccades are programmed voluntarily based on some criteria which is not yet fully understood [18,20,21,23,24,26,29].

The interest criteria attracting the eye can partially be determined by looking at the responses of neurons in the retina and visual cortices [1,2,3,4,7,11,22,49]. However, how these features are combined into a single interest criteria is unknown. Another interesting question is whether this criteria is only based on visual properties like brightness, edge content, motion, etc. In a number of recent studies it was shown that high level semantic properties of objects in the periphery were not effective in determining the next fixation point.

Two of the better-understood cognitive effects in saccade programming are *negative priming* and *inhibition of return*. Negative priming refers to a subject's reluctance to pay attention to a feature if that feature had been ignored for some reason in a previous task. Similarly inhibition of return is the inhibition of spatial locations or features which have just been fixated. Together with dead-zone and interest criteria, negative priming and inhibition of return are the elements of *covert attention*, which is defined as attentional effects of which the person has no conscious awareness [34].

1.2.1.4 Oculomotor Models

Efforts on modeling human oculomotor system mostly concentrate on the smooth pursuit system which is involved in target tracking tasks. In few studies models of saccadic system are also proposed. In order of increasing success the main approaches to the problem include linear and predicting controller models, optimization and stochastic estimation techniques and methods based on decision making [18,19,53,55].

The oculomotor system is known to be non-linear. However, modeling it as a linear time invariant system still works for small signals and some of the non-linearity can be introduced by limiting acceleration and velocity. Various aspects of linear systems approaches can be illustrated on a second order linear system offered by Westheimer in 1954 to model oculomotor behavior [18]. This second order dynamic system is causal, i.e. it cannot respond before a stimulus is presented. Its performance is the same with predictable or unpredictable stimuli. The system is also time-invariant which means that it cannot improve or change its performance in time.

It is also thought that some aspects of the oculomotor systems may be modeled by classical control theory if the observed anticipatory behavior could be represented and introduced as a fixed predictor in the signal path [18]. For example in 1963 Dallos and Jones studied the gain and phase lag of pursuit as a function of the frequency of predictable and random target motions and designed a predictor based on this data. The memory system generated an expected signal which was compared to the actual error signal. If they match the predictor was inserted in the loop, otherwise error signal was fed into the plant. Other researchers applied methods which were based on the same idea. However one major limitation of this idea was that the resulting predictor was not physically realizable. The model also failed to characterize saccadic eye motions.

In 1984 Yasui and Young developed a model of anticipation in which they seperated the two components. They assumed that the output of the system is an additive combination of pursuit and saccadic subsystems. They also proposed that the effective input to the saccadic subsystem was the positional error minus the pursuit component.

However Yasui and Young's analyses did not permit a complete evaluation of their hypothesis [18,20,21].

The models discussed so far can account for some aspects of anticipation, however they cannot explain how the system can learn to predict and they also cannot explain some apparent anticipatory effects like eye motion occuring before stimulus. The general idea in this respect is that the system responds to a hypothetical internal signal. More recent approaches which come closer to modelling such phenomena are based on optimal estimation, stochastic methods and modern control theory.

In an optimal control approach the problem is defined by the dynamics of the system, an objective function to be minimized and some constraints. In this case the dynamics of the system is determined by the mechanical properties of the eye. The objective function is task dependent. In the case of tracking the objective is to minimize the error between the target and the optical axis of the eye. Finally constraints are introduced by physiological limitations like forces applied to the system.

Kalman filter is another method used to model eye movement during tracking [18]. Kalman filter can generate predictions of target trajectory in a recursive manner. The prediction is refined at each time step by an equation of the form $x(t+1) = a(t)x(t)+b(t)y(t)$, where x is the estimate of the state of the target and y is the observed position of the target. Coefficients a and b determine the contribution of these two components based on their relative contamination by noise. In general these parameters are calculated from the linear model used for target motion and from the statistics of x and y. A serious objection to Kalman filtering is the requirement of a linear target model. It is also not considered logical to assume that the oculomotor system can identify and recall the parameters of target motion.

In general all control theoretic approaches can successfully model human oculomotor behavior in a special case and under certain assumptions. More recent models can learn to extrapolate or model target motion. However, at least the following three properties of the oculomotor system cannot be modeled by these approaches:

1. the ability to learn how to extrapolate all types of target trajectory,
2. the ability to incorporate contextual information and change behavior respectively,
3. the ability to make voluntary decisions.

The above difficulties are overcome by using decision making methods to model higher level functions of the oculomotor system [18,19,39,55,56,57]. Adaptive networks and Markov models are some approaches that are used to model decisions made during eye movements based on previous experience. A more interesting and challenging idea is to integrate these models with a stimulus based response. Research in this field is still in its infancy and applications by Jordan, Falmagne or Kowler et al. are limited to simple situations like eye motions during following a stimulus with two possible paths.

1.2.1.5 Key Properties of Biological Vision

Although our understanding of visual perception is still incomplete, in the light of the above discussion it is possible to list the key properties of biological vision as follows:

1- Fovea – Periphery Distinction: Biological vision systems process only a small part of their visual field in detail. Unlike traditional cameras used by man made imaging systems, the distribution of receptor cells on the retina is like a gaussian with a very small variance, resulting in a dramatic loss of resolution as we move away from the optical axis of the eye. The small region of highest acuity around the optical axis is called the fovea, and the rest of the retina is called periphery.

2- Overt and Covert Attention: As a consequence of fovea-periphery distinction, saccades – rapid eye movements - are used to bring images of chosen objects to fovea where resolution of fine visual detail is at its best. This physical attention mechanism is called *overt attention*. Saccadic eye movements are voluntary and require the computation of the relative position of a visual feature of interest with respect to the fovea in order to determine the direction and amplitude of the saccade. A second type of attention system called *covert attention* refers to unconscious attentional effects. These include poorly

understood complex cognitive processes, which determine the attention behavior of the system.

3- Levels of Representation: A third feature is that cells in the visual path from retina to the primary and other cortical regions respond to increasingly more complex stimuli, accompanied by larger receptive fields on the retina. For example in the primary visual cortex, simple cells respond to lines of a particular orientation, more common complex cells respond to motion, and some cells both simple and complex respond to specified corners and curvatures.

4- Serial Processing: Although the human visual system is massively parallel in structure, most visual tasks also require serial processing as the oculomotor activity results in the perception of a series of images in time. Especially in counting or comparison experiments more complex scenes lead to longer processing times in human subjects because of increased number of fixations or eye movements required to solve the task. This implies that information is collected and somehow combined after each fixation until there is enough information to make a decision.

5- Memory: Human vision also relies heavily on short and long term memory. Some cognitive effects during attention like inhibition of return or negative priming require a short-term memory mechanism. Long term memory is used to accumulate visual information during fixations and to build abstract models of the environment that can last for years.

1.2.2 Active and Attentive Robot Vision

Biological vision systems have the capability of allocating their visual resources to different parts of a scene in time by shifting their attention. This shift of attention is obtained mechanically by eye and head motions and also by higher level cognitive mechanisms in a continual loop of pre-attention and attention. The incoming stream of sub-images is then utilized to generate a spatio-temporally related sequence of features -- referred to as *attentional sequence*. The term attentional sequence is intended to convey

two important characteristics of this data: First, at each instant only a small part of the scene is attended through a fovea-fixation mechanism. Second and perhaps more fundamentally, the sequential relations between attentive behavior stress the spatio-temporal nature of the vision data. Visual understanding becomes a problem of properly interpreting the attentional sequences that are being generated when looking at an object or a scene.

Machine vision systems endowed with selective perception - motivated by biological vision - allocate their limited resources to process only the most relevant parts of the incoming data [13,35,37,38]. This is done by first implementing a retina model, where a periphery and fovea can be defined and processed at different resolutions or levels of detail. The fovea is defined to be a small region around the center of the visual field while the remaining region of the visual field is referred to as the periphery. Periphery-fovea distinction leads to a loop of pre-attentive and attentive processing as shown in Figure 1.1. In the pre-attentive stage the periphery is searched for relevant features of interest. Then a fixation on this feature is made to bring it to the fovea for detailed analysis. At each step results of fovea processing are added to a sequence of observations. The cognitive stage works with this attentional sequence in order to achieve given visual tasks. At each time step in this sequence the cognitive stage uses collected information to improve the system's knowledge and attempts to make a decision about the task being performed. If a decision can be made, the task is solved, otherwise the selective attention process continues to collect information.

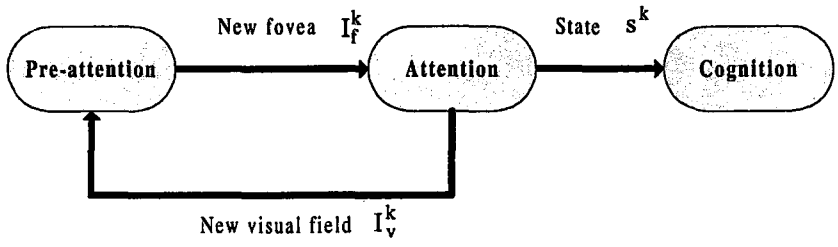

Figure 1.1 General Flow of Processing in Active Vision

Active vision research has mostly concentrated on generating fixations and controlling camera movements [12,36,44,59]. Early on, the problem of locating a fovea has been solved by data-driven saliency operators - where a sequence of camera movements

emerges from a specific image data [39,40,51]. An alternative approach based on simplified visual search mechanisms such as using attractive forces has been presented in [44,65]. A third type of mechanism based on augmented Hidden Markov models - modeling eye movements explicitly while incorporating feedback from visual cues- has been presented in [58]. A generalization of these ideas to Bayes networks and decision theory is presented in [57]. A maximum likelihood strategy for directing attention has been applied in recognition tasks [53]. Some research also focused on building electro-mechanical systems that can replicate fast and accurate saccades. Various models of attention, eye-movements and visual search were developed by both robot vision and biological vision communities [39,40,51,53,65,68,69,70]. Models of memory and internal representations and their application in active robot vision is widely discussed in both biological and computational vision literature [4,12,37,54,61,71]. Although human-like attention mechanisms in robot vision has rapidly increased, the use of visual data collected by attentive vision systems remained relatively unexplored [41,42,43,72,73,74,75,76].

1.3 Motivation

We currently lack a complete integrated theory of human attention mechanisms and oculomotor systems, although their role in visual processes is widely accepted. Since 1960s researchers were able to measure eye movements roughly and many studies concentrated on finding relationships between eye movements and the input stimulus. However these input/output relationships were never discovered. In all experiments the same stimulus could result in different responses and different stimuli could result in the same response. The only explanation, which is only now gaining support, was the effect of higher level cognitive factors in visual processing. Although cognitive factors like choice, effort, selective attention, expectations and memory were known to exist, they were always ignored or simplified. Therefore research efforts were concentrated especially on the smooth pursuit movement, rather than the saccadic system, although a few recent studies incorporated cognitive factors in very simple cases.

In active robot vision there has been a lot of work on the generation of attentional sequences either using images from active camera heads or by changing the region of interest in static images. However, the next step of linking the attentional sequences to visual tasks and the responsible higher-level mechanisms are less explored. Yet, developing models of such mechanisms prove out to be crucial to understand human vision and develop attentive robot vision systems.

Integration of visual cues in order to build a long term environment model is another visual task performed by the humans. While classical computer vision systems do not need this information, it is crucial for a mobile robot to have this capability. Current work in environment modeling is separated from active vision research and therefore the problem is not defined or solved for mobile systems with attention capability.

1.4 General Approach

Our mobile robot APES, described in section 1.7 is used to provide the basic hardware for active vision. The sensor model and selective attention mechanisms we propose in chapter 2 incorporate a novel two-camera retina model, edge based pre-attentive and attentive processing, and some of the well known covert attention effects like inhibition of return, dead zone, memory and attentive features. The general properties of attentional sequences generated by such a system are discussed in chapter 3.

Next, in chapter 4, we consider object and scene recognition tasks by an attentive system and propose two approaches regarding how to make use of attentional sequences in this problem: Markovian and evidential reasoning. The two approaches – although seemingly different from each other - have underlying common themes: First, they can be used with different pre-attentive and attentive features (color, edge, brightness, texture, etc.) without modification. Secondly, the approaches are capable of handling variations with regards to scanpaths taken. Finally, these approaches have mechanisms for learning under external supervision.

The bubble model proposed in chapter 5 is an attention based integrative memory, which is considered as a model of visual integration over fixations and a method of environment modeling for mobile robots.

1.5 Contributions of the Thesis

Major contributions of the thesis are as follows.

1. Attentional sequences are introduced as a scene modeling and recognition tool.
2. A model for visual integration over saccades, which enables vision based environment representation by active and attentive vision systems, is introduced.
3. A model of visual attention which simulates the key properties of biological vision is developed and implemented on a mobile robot – APES. New short term memory and fixation control techniques are introduced.

1.6 Outline of the Thesis

In the next subsection the mobile robot APES, used as our experimental setup throughout this thesis, is described. In section 2 we develop a model of visual attention, which simulates the key properties of visual attention. Then the output of the attention mechanism – the attenional sequence is analyzed and its representation capability is discussed. In chapter 4 two algorithms for using attentional sequences in scene recognition are developed. The bubble model for environment modeling and visual integration over saccades is proposed in chapter 5. Then these mechanisms are combined into an integrated model of active and attentive vision. We conclude by summarizing our contributions and future plans.

1.7 Experimental Setup: APES – Active Perception System

APES, shown in Figure 1.2, is a mobile robot developed in our laboratory for active vision research. Its body is driven by two conventional wheels. Using four stepping motors it can translate and rotate its body and direct its cameras to the visual stimuli by pan and tilt motions. Body rotation and camera pan axes have been designed to be co-centered, in order to simplify transformations during combined body and camera motions, and are not the same as the centerline of the cylindrical body for mechanical stability reasons.

APES is designed to be able to simulate the key properties of biological vision discussed above. Its simple hardware and flexible software libraries enable easy integration of different oculomotor and retina models as well as memory and recognition modules to build a biologically motivated vision system [75,77].

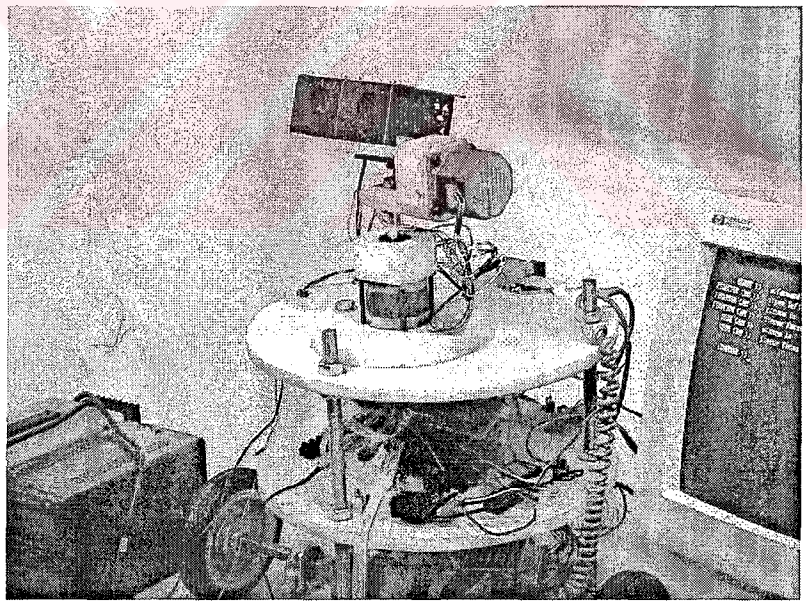

Figure 1.2 APES robot and its 2 dof camera base.

Table 1.1
Technical specifications of APES.

Height:	60cm.
Radius	37cm.
Wheel span:	52cm.
Wheel radius:	15cm.
Drive method:	Stepping motors
Power:	12 V Battery
Pan accuracy:	1.8 degrees
Tilt accuracy:	1.8 degrees
Video format:	CCIR composite
Image size:	512x512 pixels
Camera lens:	4-47 degree zoom

Figure 1.3 shows the hardware configuration of APES. The main visual processing module running on a workstation performs vision processor setup, frame grabbing, pre-attentive and attentive processing and serial communications. The on-board PC104 computer is responsible for serial communications, motor control, and camera control. All camera features including zoom angle can be controlled by the on-board computer.

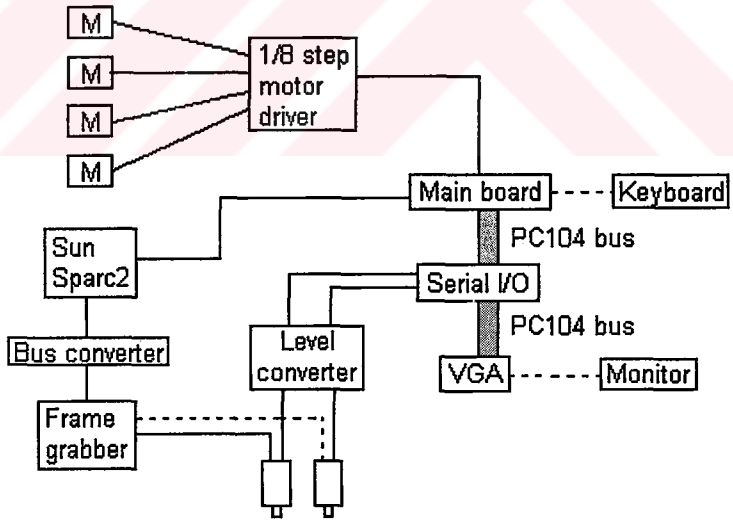

Figure 1.3 Schematic of APES

The two degrees of freedom step motor based head assembly and camera motions of APES cannot be compared to the highly developed oculomotor system. However APES can effectively control the optical axis of its camera with an accuracy of 1.8 degrees due to its step motor based drive system. Camera motions correspond to large and fast saccadic motions of the eye, which are used for fixating different spatial targets. During operation

the saccade system determines the new fixation point in the periphery and the corresponding saccade vector. This information is sent to the on-board computer which moves the camera accordingly. The new visual field is then processed by the vision system.

In addition to this physical attention mechanism, during both pre-attention and attention stages, APES can use different features to change its attention criteria and to obtain different representations of it visual environment. Currently in the pre-attentive stage APES uses either edge content (computed by the gradient) or brightness as low level attention criteria or saliency measures. In the attentive stage APES can process the fovea to extract features like edge strength, edge type, cartesian and non-cartesian primitives, relative and absolute saccade directions, etc. which give information about the nature and relative positions of features.

The ability to use different features in both stages enables APES to explore and internally represent its environment in different ways. For example by using a gradient based attention criteria and directional selectivity APES can be made to smoothly follow object contours, or by using the brightness feature it can be made to fixate on light sources, reflective objects, etc. Similarly the object whose contour is being traced can be modeled by using various attentive features. Snapshots from recognition experiments on curved metal objects, shown in Figure 1.4, demonstrate the controlled contour tracing behavior by APES. A visual field much smaller than the actual camera image is used in this case.

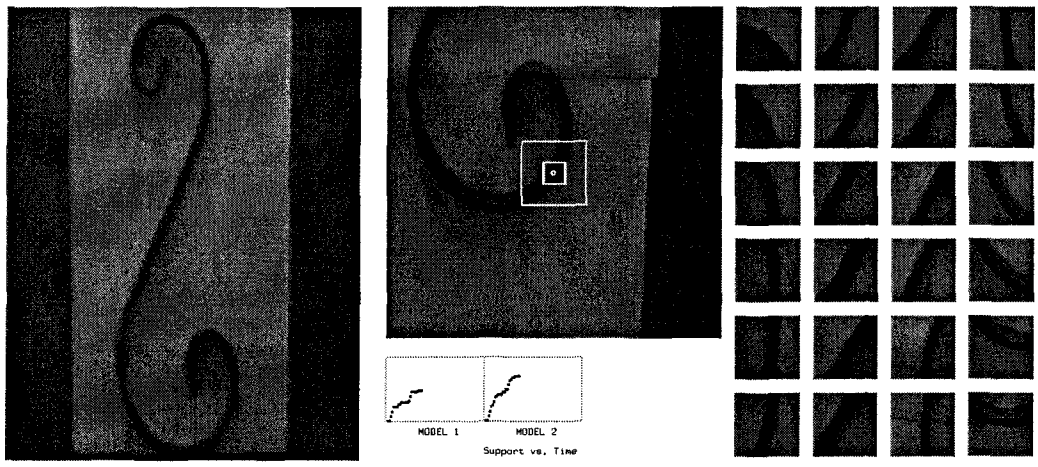

Figure 1.4 Snapshots from a recognition experiment using selective attention (right) on curved metal object (left).

In Figure 1.5 a snapshot from APES' main software is shown. The two large image boxes can display actual visual field image and its gradient or the wide and narrow angle camera images. Below the gradient image is a subsampled low resolution periphery image which is used in single camera configuration. The tiny fovea image is also shown on the left. A control window is used to select operating modes and settings, and a separate data window displays all computations, including fovea saliencies, attentive features, saccade vectors, bubble points, fixation numbers, etc.

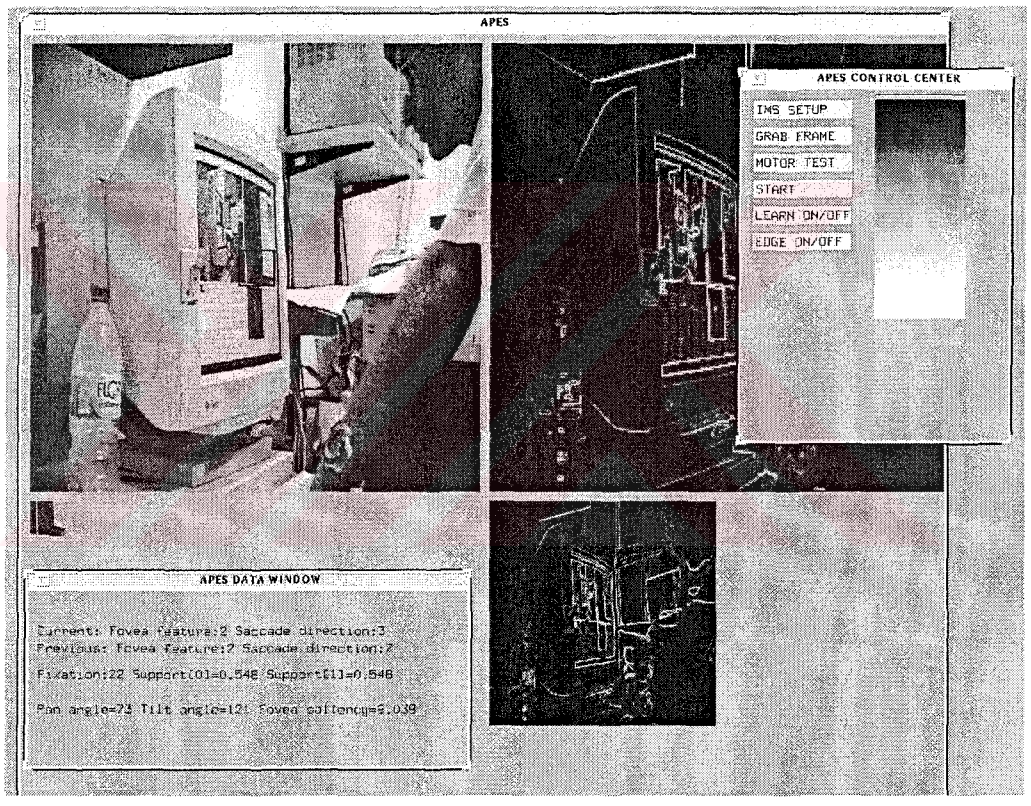

Figure 1.5 APES main software snapshot.

2. A NEW MODEL OF VISUAL ATTENTION

2.1 Pre-Attention and Attention

Overt attention behavior characterized by voluntary saccades can be modeled by a continuous loop of pre-attentive and attentive processing stages. During pre-attention, simple attentive features are computed from the periphery region in order to select the next fixation point and thus the next fovea to be fixated. Let I_v^t represent the visual field image and I_f^t represent the fovea image at time t. Let $C(I_v^t)$ denote the set of candidate foveas – determined from the visual field. For each candidate fovea $I_f^c \in C(I_v^t)$ an attention criteria $a : I_f^c \rightarrow R^+$ – a scalar valued function of interest based on the presence of simple features with low computational requirements – is computed. The candidate fovea maximizing this criteria is then selected as the next fovea:

$$I_f^{t+1} = \arg \max_{I_f^c \in C_f(I_f^t)} a(I_f^c) \tag{2.1}$$

When a selection is made, the optical axis of the camera is directed to bring that area into fovea. Such camera movements correspond to saccadic eye movements in humans. As a result, a sequence of foveas is generated. Let $I_f = (I_f^1, \ldots, I_f^T)$ be the stream of foveas looked at as of the T^{th} fixation.

In the attentive stage, each fovea I_f^t is subjected to detailed analysis in order to make an observation o^t about the state of the fovea. In general, this analysis is much more computational than the pre-attentive stage and the visual primitives that are used can be rather complex. Consider M different visual primitives and let the set of values of m^{th} visual primitive be denoted by Ω_m. The value of each visual primitive is obtained via an operator $f_m : I_f^t \rightarrow \Omega_m$ acting on the fovea I_f^t.

If Ω_m is a finite set with N_m elements, then let $\Omega_m = \{v_{m_1}, v_{m_2}, \ldots, v_{m_{N_m}}\}$ denote the set of values that f_m can take. Let Ω denote the feature space as $\Omega \stackrel{\Delta}{=} \Omega_1 \times \ldots \times \Omega_M$. Note that

$$|\Omega| = \prod_{m=1}^{M} N_m \tag{2.2}$$

Each observation $o^t \in \Omega$ then becomes a vector of visual primitive values:

$$o^t = \left[f_1[I_f^t], \ldots, f_M[I_f^t] \right] \tag{2.3}$$

Thus, as a stream of foveas $I_f = (I_f^1, \ldots, I_f^T)$ is generated, so is an attentional sequence $O^T = (o^1, \ldots, o^T)$. Hence, an attentional sequence can be visualized to be a spatio-temporally related set values of visual primitives – containing the critical visual data. Obviously, the choice of the visual primitives is of utmost importance - if we are to use attentional sequences in visual tasks.The cognition stage then operates on the observation sequence O^T in order to solve the given visual task.

2.2 Inhibition and Short Term Fixation Memory

Within this framework, visual processing consists of three basic stages of operation: pre-attention, attention, and cognition as shown in Figure 4. The visual field components are shown in Figure 5. A new fovea is found by considering overlapping candidate foveas within the visual field, computing their saliencies using an attention function $a : I_f^c \rightarrow R^+$ and designating the center of the most salient fovea as the next fixation point as explained previously. In addition, two mechanisms - inhibition and fixation memory - get activated before a saccade is made in order to avoid processing the same areas twice or going into infinite fixation loops. This is motivated by vision science findings that indicate the presence of dead zone (inhibition) and inhibition of return (fixation memory) mechanisms

that inhibit small saccades and delay fixations on an area that has just been fixated. The inhibition mechanism works as follows: An HxH pixel region I_h^t around the currently fixated fovea I_f^t, at the center of the visual field I_v^t, is defined as the inhibition region. All candidate foveas $I_f^c \in C(I_h^t)$ falling within the inhibition region are inhibited. In this manner, the inhibition mechanism controls saccade magnitudes. Although not confirmed biologically, it is also possible to define a non-uniform inhibition field and control saccade directions as shown in [43,73,75,76,77]. The memory or inhibition of return mechanism works via keeping track of previously fixated foveas and inhibiting them even if they are not within the current inhibition region. For this, we use a first-in-first-out memory $C_d = \{I_f^t, I_f^{t-1}, \ldots, I_f^{t-D}\}$ of size D. All foveas in this memory are inhibited during pre-attention. At the end of each new fixation, I_f^{t-D} is removed from while I_f^{t+1} is added to this memory.

In summary, pre-attentive processing together with inhibition and memory mechanisms are merged to form an augmented attention function $\tilde{a}: I_f^c \rightarrow R^+$ as:

$$\tilde{a}(I_f^c) = \begin{cases} 0 & if \quad I_f^c \in C(I_h^t) \\ 0 & if \quad I_f^c \in C_d \\ a(I_f^c) & if \quad I_f^c \in C(I_v^t), I_f^c \notin C(I_h^t), C_d \end{cases} \tag{2.4}$$

Note that any simple image feature as low level attention criteria can b used in the pre-attentive stage, and these criteria can be varied in order to generate fixation behaviors with different characteristics. In the experiments presented in the next sections APES uses a gradient based attention criteria, $a(I_f^c) \stackrel{\Delta}{=} \sum_{p \in I_f^c} |\nabla I(p)|$.

In the attentive stage the fixation fovea is subjected to more detailed processing. Various complex features can be extracted during attention. In general, the complexity of attentive processing is proportional to the size of the feature space Ω and the computational complexity of the features involved. For example in the object recognition experiments reported in section 4.4, a very simple feature set $\Omega = \Omega_1$ is considered. The

set Ω_1 is defined as $\Omega_1 \stackrel{\Delta}{=} \{i \mid i=0, \ldots, 7\}$ where each value i = 0,1,2,3 indicates an edge oriented i×90° and each value i = 4,5,6,7 indicates an edge oriented i×90°+45° .

Attentive processing strongly affects the performance of any further computation in the cognitive stage, where the visual task is being solved, as the feature vector strictly determines the information content of the observation sequence. For example, regardless of recognition methods being used, consider an object recognition task based on the sequence generated in the above example. The eight edge types in Ω_1 are already 45 degree rotated versions of the same edge, therefore rotation invariance can only be expected up to +/- 22.5 degrees even if edge detection is noiseless.

2.3 Attentive Features

The information content of an attentional sequence, generated as a result of active vision behavior, is limited to the information content of attentive features. Therefore, one of the issues critical to the success of vision tasks performed by an attentive system is the selection of attentive features computed from a fixation fovea. Originally the APES system used a set of edge types shown in Figure 2.1 and absolute or relative saccade directions between two successive fixations. Saccade directions are quantized into 8 chain coded directions as shown in Figure 2.2. In the case of relative direction differential chain code is used. While edge types contain visual information about the scene, saccade directions give relative spatial locations of features, which lead to rotation independent geometry.

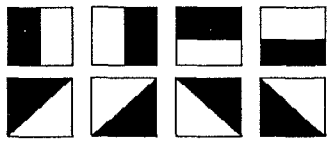

Figure 2.1 Edge types used as attentive features.

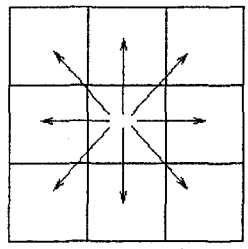

Figure 2.2 Chain coded saccade directions.

In the case of object recognition tasks discussed in section 4.4 edge types are sufficient to characterize an object's boundary. However, more complex features are required if the attentional sequence is to represent complex scenes and textures. Recordings of neural responses by Gallant et al. from Macaque area V4 and previous work on the visual cortex show that cortical cells respond to a set of Cartesian and non-Cartesian stimuli which can effectively be used to model a wide range of natural shapes and textures [2,3].

2.4 Retina Models

The attention model described above assumes fovea-periphery distinction. This is achieved by a retina model which defines the geometry of the visual field, fovea and periphery images and the corresponding processing methods. Note that the exact nature of processing which leads to the selection of a fixation point is not known. However, it is known that the peripheral image has a very low resolution compared to the fovea and the same topography is also observed in the visual cortex, where high and low level visual features are computed.

2.4.1 Geometric Model

Based on the attention mechanisms developed in the previous section, the geometry of the visual field is shown in Figure 2.3. The outermost large rectangle represents the visual field and the innermost rectangle is the fovea. The hollow rectangle between the

fovea and the boundary of the visual field is the periphery. Note that although the distribution of photoreceptors on the retina is a gaussian, it has a very small variance, which enables us to draw a strict boundary between the fovea and the periphery.

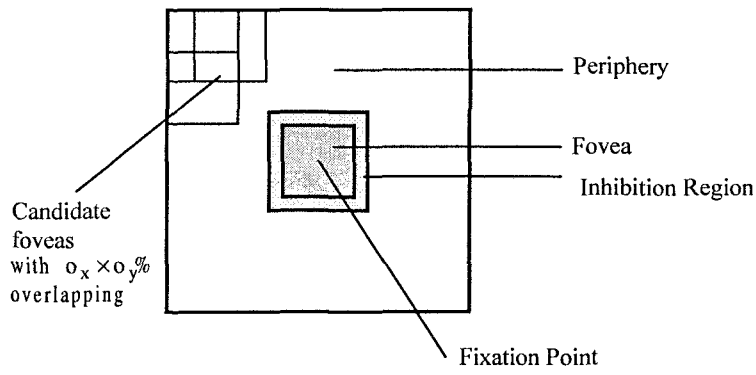

Figure 2.3 Geometric Model of the Retina

Motivated by the concept of receptive fields in the retina, we also define candidate foveas in the periphery – groups of pixels which contribute to a single saliency value. Pre-attentive processing in a candidate fovea determines whether its center will be chosen as the next fixation point by the saccade system.

Finally the inhibition region is defined to be a rectangular area around the fixation point where no pre-attentive processing is allowed. Inhibition region corresponds to the dead-zone effect, which is known to prevent very small saccades in humans. Note that the size of the inhibition region can be used to actively control saccade magnitudes.

2.4.2 Single Camera Implementation

When a single camera is used to implement the above geometric model, the non-linear characteristics of the human retina can be simulated by varying complexity of features extracted from the fovea and periphery, as described above. In the pre-attentive stage saliency of candidate fixation points are calculated by simple computations in the periphery, which is usually limited to a gradient operation. After a fixation is made, the fovea is processed to extract higher level features, as detailed in section 2.4. Note that in

the case of a single camera implementation the resolutions of the fovea and periphery images are the same.

2.4.3 Two-Camera Retina Model

In a classical active vision system the retina model is implemented using a single fixed resolution camera. In this case fovea and periphery images have the same receptor densities. As a result a wide angle visual field similar to human eye results in a very low resolution fovea image, which is not suitable for detailed feature extraction. On the other hand a high resolution fovea results in a very narrow visual field, diminishing the benefits of attentive vision. Furthermore, fovea and periphery processing have to be sequential and cannot occur simultaneously even if the required parallel processing hardware is implemented.

One alternative is a spatially variant CCD, which generates a variable resolution image similar to that generated by the retina. However the spatially variant CCD is still under development and current designs do not allow separate fovea and periphery channels for parallel processing.

A new two-camera configuration provides true variable resolution as well as parallel processing capability. To obtain a two camera fixation system periphery and fovea images must be grabbed by separate cameras which have different lens angles as shown in Figure 2.4. The wide angle camera is used to obtain a periphery image while a narrow angle camera is used for the fovea image.

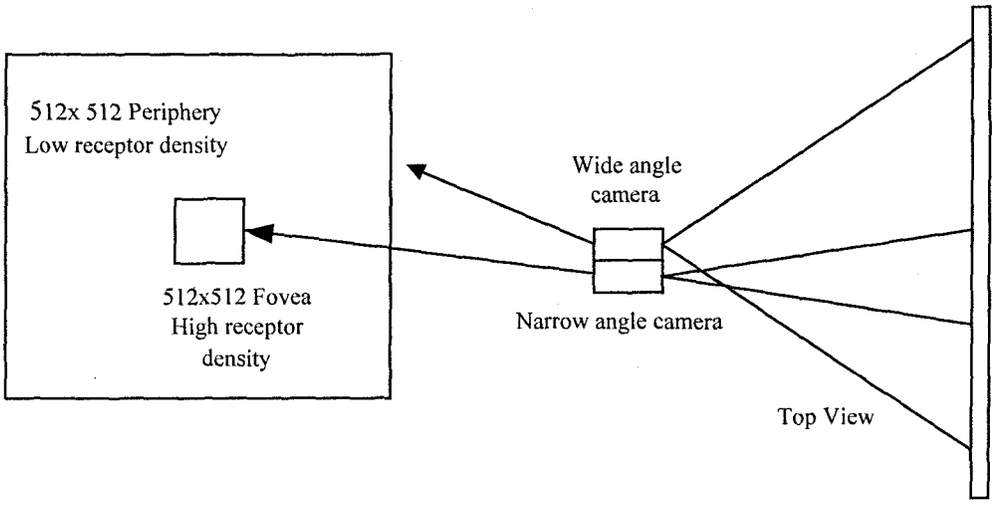

Figure 2.4 Two-Camera Retina Model.

In the APES system the periphery camera has a 46 degree wide angle lens and generates an angular pixel density of 512/46 which is around 11. The fovea camera has a 5 degree lens angle dedicating all of its 512x512 resolution to a 4 degree viewing angle and therefore obtains an angular pixel density of 512/4=128. The fovea and periphery images obtained under these conditions are shown in Figure 2.5 If a single camera system would be used for the same application, the fovea image of the same area, that would be obtained while also looking at the same periphery, is shown in Figure 2.6.

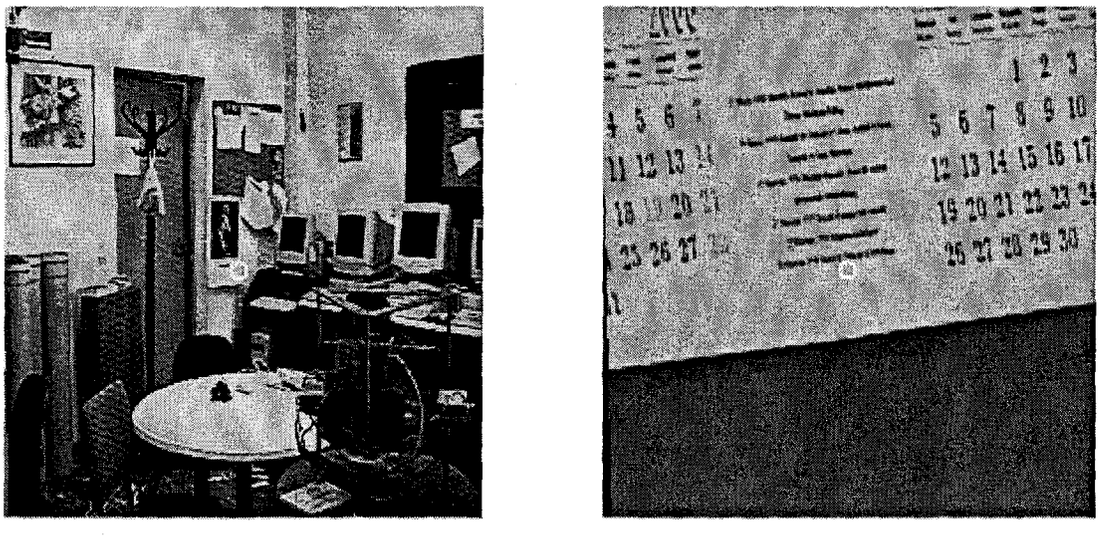

Figure 2.5 Periphery and fovea images in the two-camera retina model.

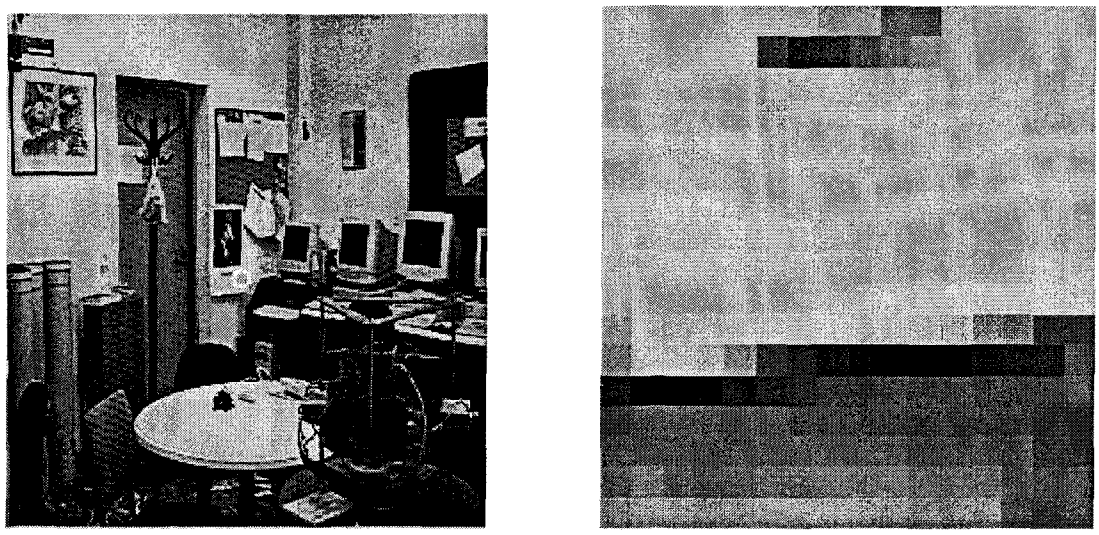

Figure 2.6 Periphery and fovea images in a single camera or stereo system.

The two cameras are fixed together such that their optical axes are parallel and as close as possible. There is a horizontal separation of about 5 centimeters between optical axes which results in a foveal image which is not exactly at the center of the peripheral image. This error is corrected in software. The controllable zoom lens cameras of APES also enable to change the areas covered by periphery and fovea images and therefore the pixel densities used. This can be done dynamically depending on the requirements of the task.

When an attentive system uses a two-camera configuration the large periphery image is searched for a fixation point and when it is found this point is converted to fovea camera coordinates and the amount of motion required for fixation is calculated. Then the camera assembly is directed to the new fixation point and a fovea image is grabbed. While this image is being analyzed for higher level features, the periphery camera can be used to search for the next fixation point if parallel processing is possible. Figure 2.7 shows the pre-attention - attention loop in the case of a two camera configuration.

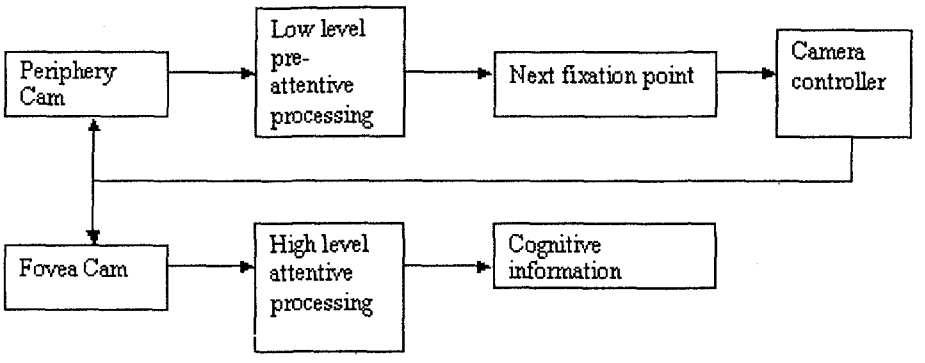

Figure 2.7 Attention procedure in the two camera setup.

3. THE ATTENTIONAL SEQUENCE

By definition, regardless of its hardware, software or biological structure, if a vision system is attentive, in the sense that it collects information from different spatial targets at different times, this system must be generating attentional sequences. Based on this common property, we ask the following question: Given a single attentive system with deterministic attention mechanisms and two or more visual scenes, is it possible to understand which scene is being viewed by looking at the attentional sequence generated by the system.

3.1 The Sequence Space

Given M different features which can take values $\Omega_m = \{v_{m_1}, v_{m_2}, \ldots, v_{m_{N_m}}\}$ as described in section 2.1 the feature space is $\Omega \stackrel{\Delta}{=} \Omega_1 \times \ldots \times \Omega_M$. Then the total number of different sequences of length T is given by $T^{|\Omega|}$. These sequences form the *sequence space*. For example in the simplest single feature case the feature space is one dimensional. If this feature is allowed to take 4 different values, then the sequence space consists of 4^4=256 different sequences. An active vision system uses information from this space to solve any given classification task.

In a two object recognition task, similar to those we have been working on, the objective is to divide the sequence space into 4 partitions: 1) sequences which can be generated only when viewing the first object, 2) sequences which can be generated only when viewing the second object, 3) sequences that cannot be generated while viewing any of the two objects, 4) sequences which can be generated while viewing both objects (intersection region). Note that this partitioning of the sequence space may be simple or complex and distributed as illustrated in Figure 3.1.

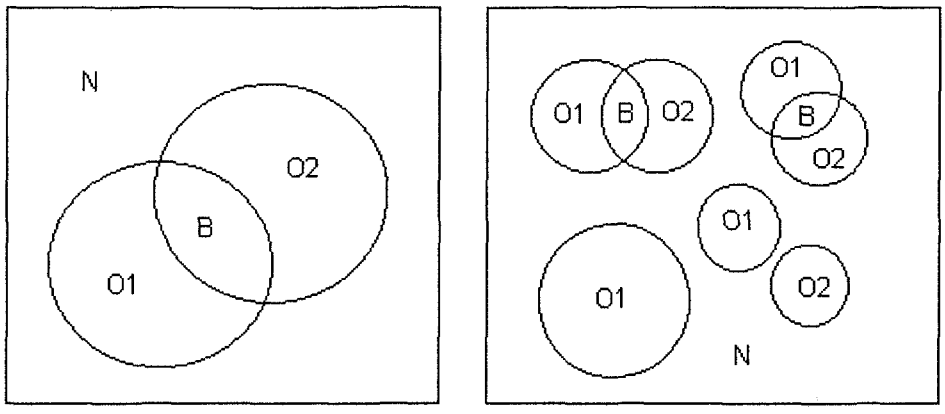

Figure 3.1 Illustrations of simple and complex partitions of the sequence space.

Interestingly, we found out that even for a very small sequence space consisting of 256 sequences, there are 4^{256} possible partitions that can occur in a 2 object classification task. This means that 4^{256} different pairs of objects can theoretically be classified by such a simple attentive vision system. Although this theoretical result is unlikely to be true in practice it clearly demonstrates the power of sequence based classification and attentive vision systems.

3.2 Is the Sequence Space Partitioned?

The existence of some kind of simple or complex partitioning in the sequence space is required for any attentive classification algorithm to be successful. This can be shown experimentally. For this purpose we generate the sequence space for a real world task using our robotic system. In more than 200 experiments performed on three objects we generated attentional sequences consisting of 4 and 5 fixations. The experiments continued until almost all possible sequences were generated on the given objects and no new sequence could practically be observed. We found out that the sequences generated on these three objects, which were 2D shapes "4", "5" and "8" with similar edge features, were mostly different. The region of intersection mentioned above was very small and had only 1 or 2 sequences among 50. In total only a small part of the space was used by these objects, but more than 90% of the sequences corresponding to at least one object belonged

to a single object. Sequences generated in the experiments are shown in Table 3.1 below. Sequences which belong to the intersection regions are in bold. A set representation of the sequence space for these experiments is also shown in Figure 3.2.

Table 3.1
50 sequences of five fixations on 2D shapes "5", "8", and "4".

2 2 1 2 2	**3 3 3 3 3**	2 3 3 3 3
3 3 3 3 3	1 1 3 3 3	2 2 1 2 2
2 2 3 3 0	3 3 1 3 0	3 2 1 2 2
2 3 3 2 3	**0 0 3 0 3**	3 2 2 2 2
2 3 0 3 0	0 3 0 3 2	3 2 2 2 3
1 3 1 3 0	3 3 3 3 1	2 2 3 0 3
3 3 2 3 3	2 3 1 0 2	**3 2 3 3 3**
3 3 3 0 3	3 3 1 3 3	2 2 0 3 2
1 3 3 3 2	**3 3 3 3 0**	**0 3 3 2 3**
2 3 3 0 2	2 3 3 3 3	2 2 2 1 3
1 2 3 3 3	3 3 3 3 3	3 2 1 2 0
3 0 3 0 1	3 0 3 3 2	3 2 3 2 1
2 0 3 1 2	**3 3 3 3 0**	3 2 2 3 3
2 3 3 2 3	3 2 3 0 0	2 2 2 1 2
2 3 2 0 2	**3 2 3 3 0**	2 3 1 1 2
3 3 3 3 0	**3 3 3 3 3**	2 2 2 3 1
3 2 3 1 1	3 3 3 2 3	2 2 2 2 1
3 3 3 3 1	1 0 3 0 3	2 2 2 3 1
3 2 0 2 2	**3 2 3 3 3**	2 2 2 2 0
2 2 0 3 0	3 3 2 2 2	3 2 2 1 2
1 3 3 3 1	1 3 3 2 3	3 2 3 2 0
2 3 1 2 0	**2 2 3 3 3**	3 3 3 1 2
1 2 3 3 0	**3 3 3 3 3**	2 2 0 2 3
3 3 2 0 3	1 2 0 3 3	2 2 2 0 0
3 1 3 2 1	3 3 3 2 3	0 3 2 0 2
3 2 1 2 3	1 3 3 3 3	2 2 2 0 2
3 3 3 2 1	0 3 3 3 2	3 2 2 1 3
3 2 3 3 0	3 2 3 3 2	0 3 3 2 2
1 2 3 2 2	**3 3 2 3 3**	1 2 2 0 3
0 2 3 0 0	0 0 2 0 2	3 2 2 2 0
2 2 3 3 3	3 3 2 2 3	2 2 3 3 2
0 3 3 2 3	3 3 3 2 3	1 2 2 2 3
1 3 3 1 3	3 1 3 3 2	3 2 2 3 3
2 0 3 0 3	2 2 1 6 3	0 2 2 2 1
2 0 3 3 3	1 3 0 3 2	3 3 2 2 3
3 0 3 0 0	0 3 2 3 1	1 3 2 3 1
3 2 0 2 2	3 2 3 1 2	2 1 3 3 1

0	**0**	**3**	**0**	**3**		2	3	0	3	3		0	3	2	2	2
0	2	1	0	1		0	3	2	3	2		3	0	2	1	2
3	3	0	3	0		3	3	3	2	3		3	0	2	3	2
0	2	1	1	3		0	3	3	3	3		2	3	2	2	3
0	3	3	3	1		**3**	**2**	**3**	**3**	**3**		1	2	2	2	3
2	3	2	0	3		0	3	0	3	2		3	3	0	3	3
2	3	2	0	2		1	3	0	3	3		2	0	3	3	1
2	2	0	3	0		3	2	1	3	2		2	2	0	2	0
3	3	2	1	2		3	1	0	3	2		3	3	2	2	3
2	3	2	2	0		1	3	1	0	3		**3**	**3**	**2**	**3**	**3**
3	0	3	0	2		1	3	2	3	3		**3**	**3**	**2**	**0**	**3**
3	1	0	2	3		**0**	**3**	**3**	**2**	**3**		3	2	2	2	2
1	3	1	2	3		3	2	3	2	1		0	3	2	0	3

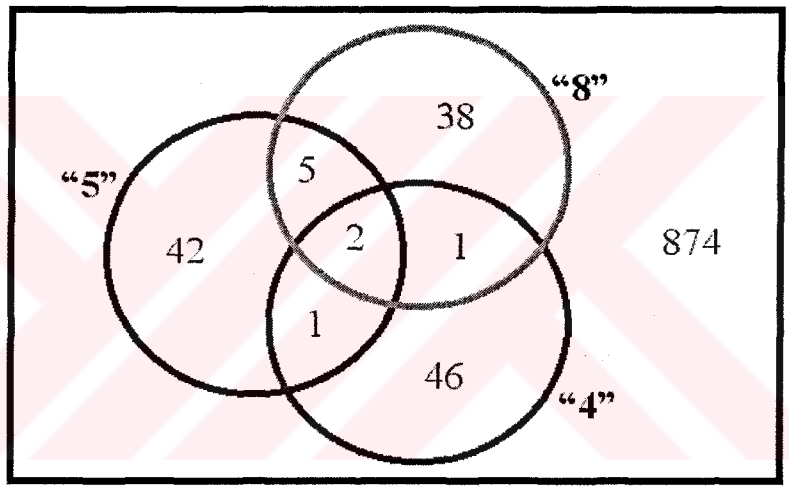

Figure 3.2 Experimental partitions in the sequence space for 2D shapes.

These results showed that the sequence space was partitioned as we expected and also that the intersection region was small, theoretically enabling the classification of scenes even when they have similar features. However, it is not possible to understand the geometry of this space, as we do not have a distance measure for attentional sequences. Therefore, whether this partitioning could be detected by a particular classification method is a different problem discussed in the next section.

4. SCENE RECOGNITION BY AN ATTENTIVE SYSTEM

An attentive vision system - unlike classical computer vision systems – requires new approaches to make use of the attentional sequences generated as a result of selective attention behavior. In this section we propose two new approaches for modeling and recognition of attentional sequences.

The visual task is defined as follows: Suppose the vision system is looking at a scene in an attentive manner and thus generating an attentional sequence O^T. Furthermore, suppose that the system knows about L different scenes - to which the scene currently being looked could belong or not. Then find $l^* \in L$ that best explains the observed attentional sequence O^T.

4.1 Markov Models and Reasoning

In this approach the attentional sequence $O^T = (o^1,...,o^T)$ is considered as a discrete Markov process with an alphabet Ω [78]. This process is associated with the transition probability matrix A of dimension $|\Omega| \times |\Omega|$.

$$A = \{P(o^{t+1} = v^j | o^t = v^i)\} = \{a_{ij}\} \quad where \quad v^i, v^j \in \Omega \text{ and } \sum_{j \in \Omega} a_{ij} = 1 \quad , \forall i \in \Omega \qquad (4.1)$$

Here $P(o^{t+1} = v^j \mid o^t = v^i) = a_{ij}$ denotes the probability of getting a feature value v^j after having observed v^i. In a Markov process, each observation o^t at time t is called a *state*. In our case, each observation o^t represents the state of the fovea with respect to the attentive features.

The transition probability matrix is a probabilistic model of expected fixation sequences that can be generated while looking at an object. Thus, if we have a library of L object or scene, each can be represented by a different transition probability matrix A^l.

These matrices are learned after looking at these objects or scenes in a repeated manner – based on the attentional sequences generated. The learning procedure is explained in detail in Section 3.3.

When presented with a new object or a scene, the system starts looking at it and an attentional sequence O^T emerges. Let $P(o^{t+1} \mid o^t, l)$ denote the probability of observing o^{t+1} after having observed o^t with the transition probability matrix A^l. The conditional observation probability $P(O^T \mid l)$ of this sequence by model l is given by:

$$P(O^T|l) = P(o^1) \cdot \prod_{t=1}^{T-1} P(o^{t+1}|o^t, l) \quad \text{where} \quad P(o^1) = \frac{1}{|\Omega|} \tag{4.2}$$

Hence, the correct classification l^* of an unknown scene can then designated as the the library model $l \in L$ maximizing $P(O^T \mid l)$

$$l^* = \arg \max_{l \in L} P(O^T|l) \tag{4.3}$$

It must be noted that as more information is collected and thus the attentional sequence becomes longer, the value of $P(O^T \mid l)$ decreases and must therefore be scaled accordingly [78].

4.2 Evidential Models and Reasoning

In this approach the attentional sequence $O^T = (o^1, \ldots, o^T)$ is considered as a sequenced body of evidence – which can then be used to support competing propositions concerning the correct classification of a scene to different degrees [76,79]. The basic idea is to use a number between zero and one to indicate the degree of support a body of evidence provides for each proposition. Different bodies of evidence are then combined to find the proposition which is most supported.

Let l^* be the correct classification of the scene. Suppose the set of its possible values are given by L - the frame of discernment. Then propositions of interest are precisely those of the form "the true value of l^* is in A" where and hence are in one to one ` orrespondence with the subsets 2^L of L. Thus, we use $A \in 2^L$ to denote a proposition. In classification, we are in particular interested in propositions of the form:

$$A_l = \{\ l\ \},\ l = 1,\ldots,L \quad where \quad L = |L| \tag{4.4}$$

Now suppose for each proposition A_l, we have a transition frequency matrix $T_l : \Omega \times \Omega \rightarrow [0,\infty]$. Each entry $T_l(v^i, v^j)$ represents the weight of evidence attested to observing v^j after having observed v^i.

Now let $o^t \in \Omega$ be an observation at time t. This observation attests evidence for each proposition A_l. Let $\omega : 2^L \times \Omega \rightarrow [0,\infty]$ represent the weight of evidence function. Then,

$$\omega(A_l, o^t) = T_l(o^{t-1}, o^t) \tag{4.5}$$

In evidential reasoning, the degrees of support for various propositions discerned by L is determined by the weights of evidence attesting to these propositions. Let $s_l : 2^L \times \Omega \rightarrow [0,1]$ define a simple support function focused on A_l. Then s_l can be defined as

$$s_l(A, o^t) = \begin{cases} 0 & if \quad A_l \not\subset A \\ s_l(A_l, o^t) & if \quad A_l \subset A,\ A \neq L \\ 1 & if \quad A = L \end{cases} \tag{4.6}$$

where $s_l(A_l, o^t) = 1 - e^{-c\omega(A_l, o^t)}$.

Note that s_l is a belief function with basic probability number $m(A_l)=s_l(A_l, o^t)$, $m(L)=1-s_l(A_l, o^t)$, $m(A, o^t)=0$ for all other $A \subset L$ that does not contain A_l.

However, each evidence points to a set of propositions A_l , l=1,...,L with different degrees of support $s_l(A_l, o^t)$. Since $A_l \cap A_k = \varnothing$ each proposition conflicts with the other. Hence the effect of each is diminished by the other. The orthogonal sum $s_l^i : 2^L \times \Omega \rightarrow [0,1]$ of the simple support functions s_i focused on A_l are given with basic probability numbers

$$m(A_l, o^t) = \frac{s_l(A_l, o^t) \prod_{\substack{i=1 \\ i \neq l}}^{L} (1 - s_i(A_i, o^t))}{1 - \prod_{i=1}^{L} s_i(A_i, o^t)} \quad \text{and} \quad m(L, o^t) = \frac{\prod_{i=1}^{L} (1 - s_i(A_i, o^t))}{1 - \prod_{i=1}^{L} s_i(A_i, o^t)} \tag{4.7}$$

and

$$s_l^i(C, o^t) = \begin{cases} 0 & \text{if } C \text{ contains none of } A_l,\ l = 1,\ldots,L \\ \dfrac{s_l(A_l, o^t) \prod_{\substack{i=1 \\ i \neq l}}^{L} (1 - s_i(A_i, o^t))}{1 - \prod_{i=1}^{L} s_i(A_i, o^t)} & \text{if } C \text{ contains } A_l \text{ but not } A_i,\ i = 1,\ldots,L,\ i \neq l \\ \dfrac{\sum_{C \cap L} s_k(A_k, o^t) \prod_{\substack{i=1 \\ i \neq k}}^{L} (1 - s_i(A_i, o^t))}{1 - \prod_{i=1}^{L} s_i(A_i, o^t)} & \text{if } C \text{ contains some of } A_l,\ C \neq L \\ 1 & \text{if } C = L \end{cases} \tag{4.8}$$

The effect of s_l^i is to provide instantaneous support based on conflicting heterogeneous evidence for each proposition A_l. The calculation of instantaneous support for a two hypotheses case is schematically shown in Figure 4.1.

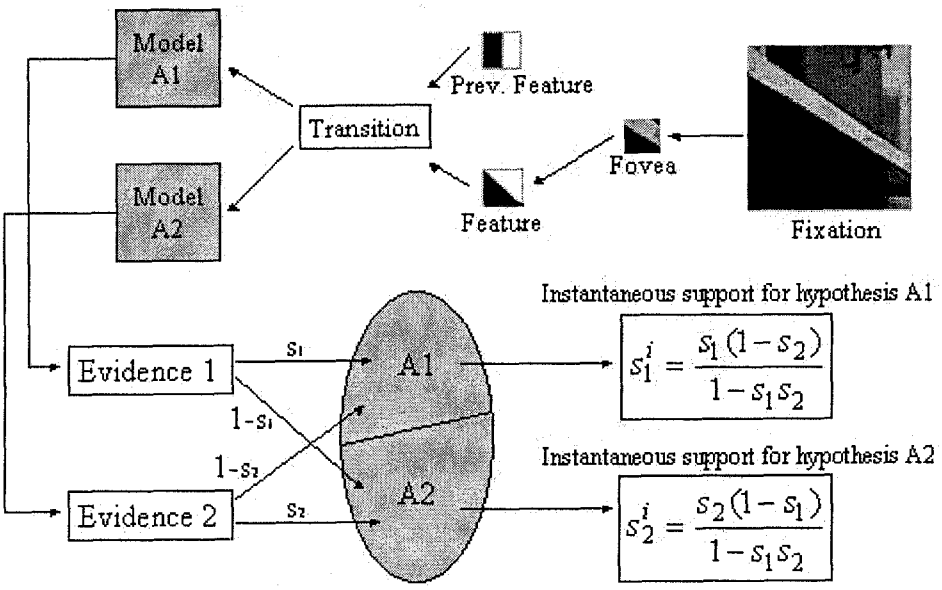

Figure 4.1 Calculation of instantaneous support for a two hypotheses case.

In order to find the total support s_l^t for each proposition A_l, the so-far total cumulated support has to be combined with the instantaneous support s_l^i. This is the case of homogeneous evidence - evidence strictly supporting a single proposition.

Let $s_l^t : 2^L \times \Omega^t \to [0,1]$ denote the cumulative support function for an attentional sequence O^t. Suppose a new fixation is made and observation o^{t+1} is made. Based on the evidence provided by this observation, instantaneous evidence $s_l^i(A_l)$ is generated for each proposition A_l. Bernoulli's rule of combination provides a reasonable way of combining s_l^i focused on Al with $s_l^i(A_l)$ and s_l^t focused on A_l with $s_l^t(A_l)$. The cumulative support $s_l^{t+1} : 2^L \times \Omega^{t+1} \to [0,1]$ is defined recursively as the orthogonal sum $s_l^{t+1} = s_l^t \oplus s_l^i$:

$$s_l^{t+1}(C, O^{t+1}) = \begin{cases} 0 & \text{if } C \text{ does not contain } A_l \\ 1-(1-s_l^i(A_l, o^{t+1}))(1-s_l^t(A_l, O^t)) & \text{if } C \text{ contains } A_l \\ 1 & \text{if } C = L \end{cases} \quad (4.9)$$

Then the result of classification is given by,

$$l^* = \arg\max_{l \in L} s_l^{t+1}(A_l, O^{t+1}) \tag{4.10}$$

The combined total supports are checked at the end of each fixation to find a proposition supported sufficiently higher than the others. The scene corresponding to this proposition is selected as describing the current scene best. The calculation of temporal support for a two hypotheses case is schematically shown in Figure 4.2.

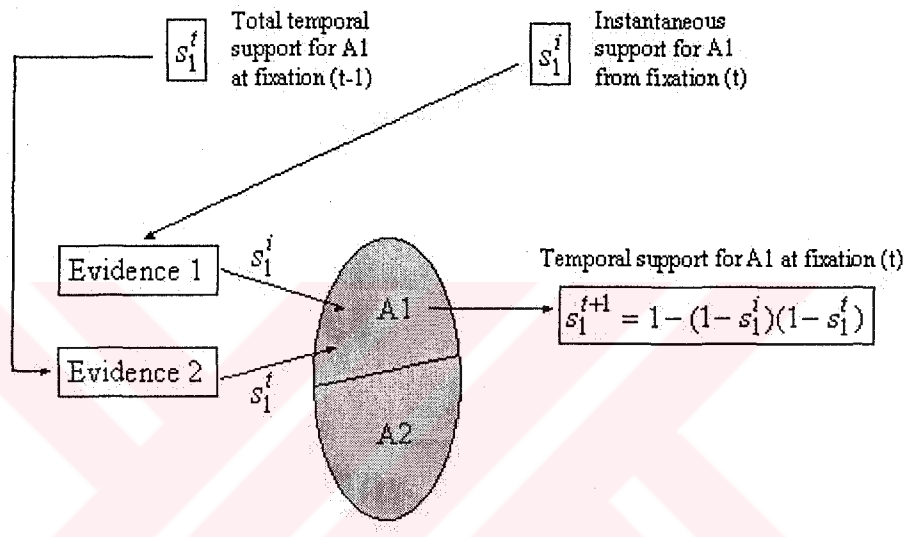

Figure 4.2 Calculation of temporal support for a two hypotheses case.

4.3 Learning Scene Models

In creating a model for each scene $l \subseteq L$, which may correspond to an object image or a complex scene, the robot starts observing the scene in an attentive manner. As it is consecutively fixating and forming observations, the transition $T_l(o^{t-1}, o^t)$ between two consecutive observations in this scanpath is recorded by incrementing the frequency of that particular transition by 1. Hence, for any library model, the number of transitions between any pair of feature vectors forms a $|\Omega| \times |\Omega|$ matrix. In the Markov approach, these matrices are converted into transition probabilities by normalizing them row by row and adding a small offset value to cope with non-existing transitions. In evidential reasoning, these matrices serve directly as weights of evidence. The modeling stage is

critical to performance of the two approaches in recognition. To obtain a perfect model all parts of a scene must be observed equally during learning fixations. Therefore, the learning period as determined by the length of the attentional sequence must be long enough to allow different scanpaths to be taken. A partial model that does not include all possible scanpaths and thus all possible feature transitions will mean that the scene is incompletely modeled.

4.4 Experiments

In order to study the efficacy of attentional sequence based recognition, APES has taken part in more than 500 experiments. Our aims in these experiments are as follows: 1.) Demonstrate the performance of Markov and evidential reasoning as sequence classification methods using simple and complex scenes; 2.) Study how variations in the learning period – the length of the attentional sequences used for learning affect the performance; 3) Understand the effects of modelling on classification performance.

In these experiments APES used a 200x200 pixel visual field and a 40x40 pixel fovea. The overlap between candidate foveas was 50% and a fixation memory depth D=10 is used to inhibit the last 10 fixated foveas. The pre-attentive attention criterion for each candidate fovea I_f^c is $\sum_{p \in I_f^c} |\nabla I(p)|$. Inhibition and memory mechanisms are employed to form the attention function as explained in Section 2.2. In the attentive stage the feature space consists of $\Omega = \Omega_1$ corresponding to 8 different orientations of a simple edge feature computed by the operator $f_1 = \arg\max_{i \in \Omega_1} S_i(I_f^t)$ where $S_i(I_f^t)$ is the 3x3 operator for detecting edges with an orientation of i degrees. In these experiments selection of simple pre-attentive and attentive features is intended to remove ambiguity in feature extraction stages and understand the exact capability of an attentional sequence as a tool for object recognition and scene classification. All experiments are performed under normal lighting conditions with both ceiling mounted fluorescents and daylight from windows. Typically, two fixation sequences generated by our robot while looking at the same scene are never identical even if there is no variation in the scene. This is caused by 1.) Slight variations in

the first fixation point; 2.) Small positioning errors in the camera head assembly; 3.) Frame grabber noise; 4.) Variations in lighting conditions. Even a one pixel wide difference in the fixation point can lead to a new visual field image for the next fixation, which results in a completely different attentional sequence as fixations proceed.

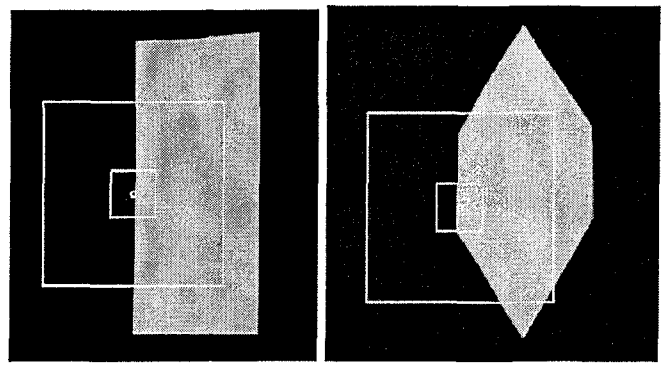

Figure 4.3 Simple scenes containing rectangle and polygon.

4.4.1 Simple Scenes

The first set of experiments was performed on simple 2D shapes hanging on a black background as shown in Figure 4.3. The system is expected to decide which scene is being viewed by analyzing the generated sequences using the Markov and evidential reasoning methods developed above. The shapes are chosen such that scene 1 of a rectangle, contains only horizontal and vertical edges, while scene 2 of polygon, contains only two vertical edges and more diagonal edges.

Table 4.1
Scene 1 - Learning using sequences of length 10.

	0	1	2	3	4	5	6	7
0	1	0	0	1	0	0	0	0
1	0	0	0	2	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	1	1	2	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Table 4.2
Scene 2 – Learning using sequences of length 10.

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	1
4	0	1	0	0	1	0	0	0
5	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	1	0	0	1

In the first set of experiments scenes 1 and 2 are used. Learning is based on attentional sequences of length 10. The observed feature transition frequencies are shown

in Table 4.1 and Table 4.2. Even with attentional sequences of length 10, these matrices start to become differentiable. The matrix for scene 1 favors no transitions between diagonal features 4, 5, 6, and 7, as compared to that of scene 2. For recognition experiments, 20 experiments with attentional sequences of length 10 are conducted. Figure 4.4 and Figure 4.5 show the generated sequences O^{10} and recognition results for both approaches. Probability values for the Markov approach are given on a log scale. Using as low as 10 fixations during both learning and classification, different feature sequences can be recognized as belonging to the correct shape with a fairly good rate.

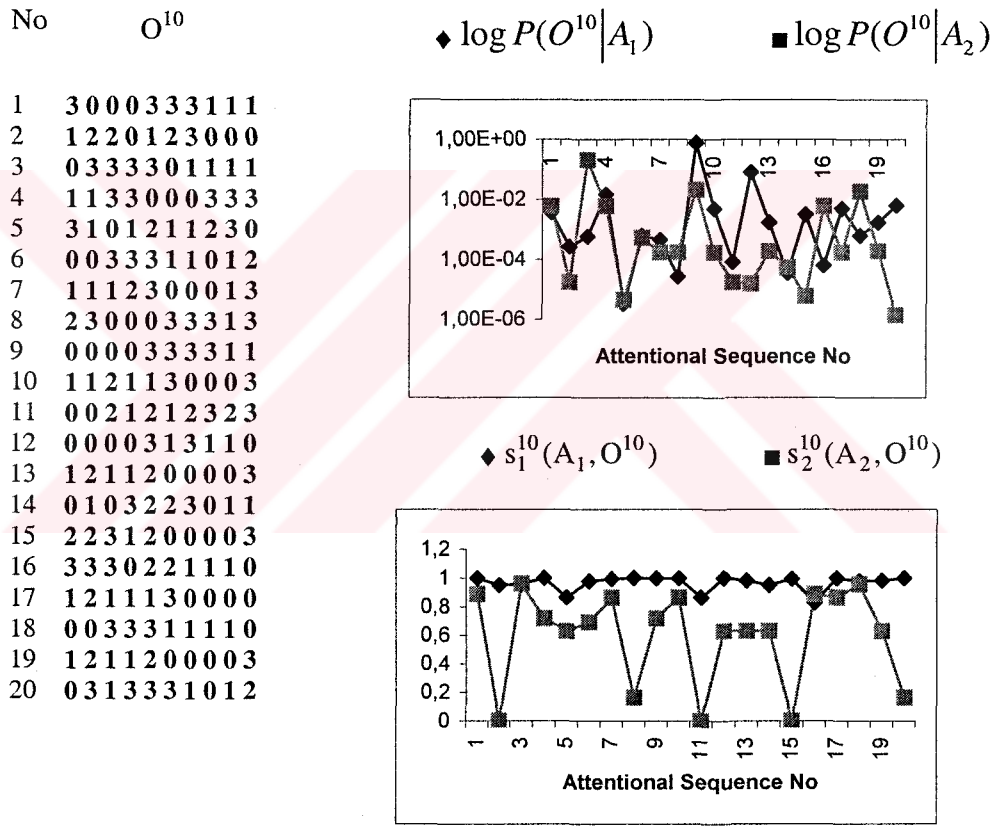

Figure 4.4 Results after 10 fixations on Scene 1 with 10 fixation learning on Scene 1 and Scene 2. Recognition rate is 65% with Markov models and 90% with evidential reasoning.

No	O^{10}
1	7021135374
2	1111111177
3	7142111133
4	1161111473
5	1273311112
6	1177331111
7	1313711122
8	1477331111
9	3111121117
10	0142111511
11	1333371173
12	7702333371
13	4211111121
14	1077331111
15	1121411333
16	7012311112
17	1147733511
18	1012111117
19	4117331161
20	1112214003

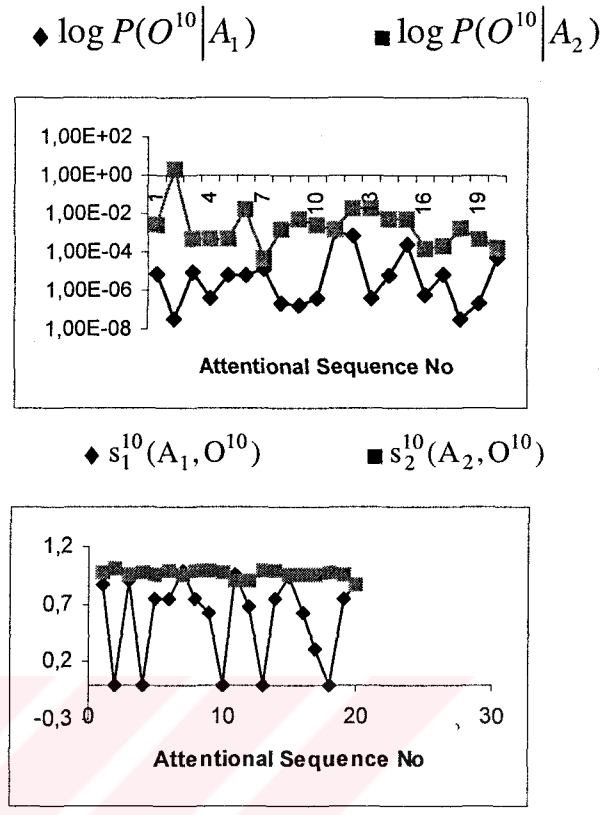

Figure 4.5 Results after 10 fixations on Scene 2 with 10 fixation learning on Scene 1 and Scene 2. Recognition rate is 100% with Markov models and 90% with evidential reasoning.

Note that the fixation camera is not following a pre-defined boundary or trajectory, therefore the 20 sequences generated during these experiments are completely different. Our classification methods are sensitive to favored transitions in the sequences based on the apriori generated models. Sequences, which include these highly favored transitions, are immediately recognized with a high margin. Others which do not include them are either incorrectly classified or return only a slightly better result compared to the competing model. Another reason for incorrect classification is the possibility of generating very similar or even identical sequences on two different scenes. However, correct classification rates indicate that this intersection region is small, and both methods work.

In the next set of experiments, we increased the learning period to 30 fixations. Differences between the two shapes are expected to become more announced. However, as observed in feature frequency matrices in Table 4.3 and Table 4.4, this may not be the case. The discriminating transitions 4, 5, 6, and 7 between Scene 1 and Scene 3 were better modeled in the previous 10 fixation models. This result shows that increasing learning

sequence size does not necessarily lead to better models and improved recognition performance due to the above-mentioned variations in sequences.

Table 4.3
Scene 1 - 30 fixation learning

	0	1	2	3	4	5	6	7
0	5	2	0	1	0	0	0	0
1	2	7	1	3	0	0	0	0
2	1	0	0	0	0	0	0	0
3	1	3	0	3	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Table 4.4
Scene 2 - 30 fixation learning

	0	1	2	3	4	5	6	7
0	0	1	1	0	1	0	0	0
1	1	6	0	3	0	0	0	0
2	0	2	2	0	0	0	0	0
3	0	0	0	3	0	0	0	3
4	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0
7	2	1	0	0	0	0	0	1

No	O^{10}
1	3 0 0 0 3 3 3 1 1 1
2	1 2 2 0 1 2 3 0 0 0
3	0 3 3 3 3 0 1 1 1 1
4	1 1 3 3 0 0 0 3 3 3
5	3 1 0 1 2 1 1 2 3 0
6	0 0 3 3 3 1 1 0 1 2
7	1 1 1 2 3 0 0 0 1 3
8	2 3 0 0 0 3 3 3 1 3
9	1 1 1 1 2 2 2 1 1 2
10	0 0 0 0 3 3 3 3 1 1
11	1 1 2 1 1 3 0 0 0 3
12	0 0 2 1 2 1 2 3 2 3
13	0 0 0 0 3 1 3 1 1 0
14	1 2 1 1 2 0 0 0 0 3
15	0 1 0 3 2 2 3 0 1 1
16	2 2 3 1 2 0 0 0 0 3
17	3 3 3 0 2 2 1 1 1 0
18	1 2 1 1 1 1 3 0 0 0
19	0 0 3 3 3 1 1 1 1 0
20	1 2 1 1 2 0 0 0 0 3

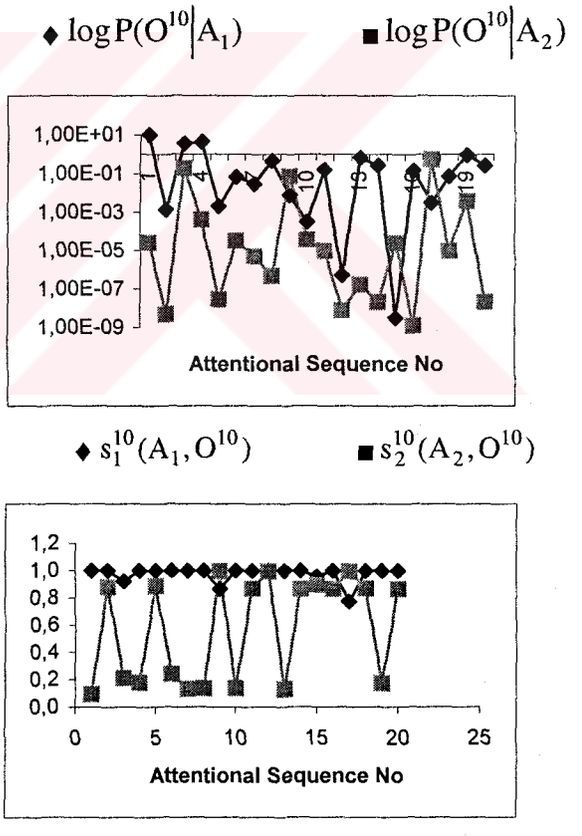

Figure 4.6 Results after 10 fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 85% with Markov models and 90% with evidential reasoning.

Results of recognition experiments using models learned from 30 fixations for Scene 1 and Scene 2 are shown in Figure 4.6 and Figure 4.7. Although an improvement in modeling and classification perfomance cannot be guaranteed by increasing the learning period, an improvement in consistency of results is observed in these results. For example

in Figure 4.7 we had significantly bad results in experiments 11-14 with both methods. Also in Figure 4.6, where recognition rate was good, both methods returned wrong results in the same 2 experiments out of 20. The remaining 1 sequence, which could not be classified correctly by Markov models, was classified correctly by support functions only by a very small margin.

For the last set of experiments a learning sequence size of 50 is used. Table 4.5 and Table 4.6 and Figure 4.8 list models generated by a 50 fixation learning run. Once again the diagonal edges of Object 2 are poorly modeled. Recognition results are shown in Figure 4.8 and Figure 4.9. Results for Object 1 are 100% correct as its model dominates over Object 2 even more than in 30 fixation models. Sequences from Object 2 are poorly recognized with the same rates as before. Consistency of results using the two approaches are again very good and in general much better than experiments with 10 fixation learning.

No	O^{10}
1	7 0 2 1 1 3 5 3 7 4
2	0 1 3 3 3 1 1 1 1 1
3	1 1 1 1 1 1 1 1 7 7
4	7 1 4 2 1 1 1 1 3 3
5	1 1 6 1 1 1 1 4 7 3
6	1 2 7 3 3 1 1 1 1 2
7	1 1 7 7 3 3 1 1 1 1
8	1 3 1 3 7 1 1 1 2 2
9	1 4 7 7 3 3 1 1 1 1
10	3 1 1 1 1 2 1 1 1 7
11	0 1 4 2 1 1 1 5 1 1
12	1 3 3 3 3 7 1 1 7 3
13	7 7 0 2 3 3 3 3 7 1
14	4 2 1 1 1 1 1 1 2 1
15	1 0 7 7 3 3 1 1 1 1
16	1 2 1 4 1 1 3 3 3 7
17	0 1 2 3 1 1 1 1 2 1
18	1 4 7 7 3 3 5 1 1 1
19	0 1 2 1 1 1 1 1 7 4

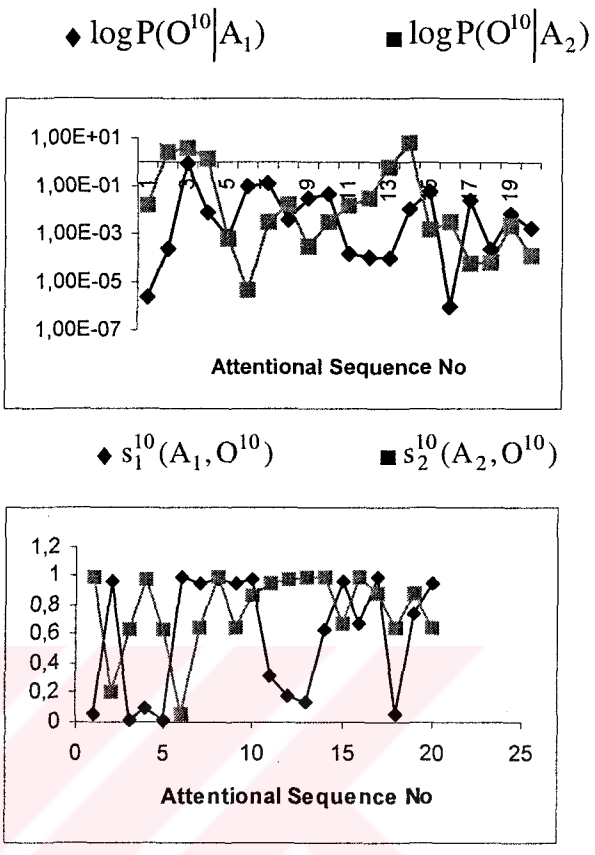

Figure 4.7 Results after 10 fixations on Scene 2 after 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 50% with Markov models and 60% with evidential reasoning.

Table 4.5
Object 1 - 50 fixation learning

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	9	2	1	2	0	0	0	0
1	1	9	1	6	0	0	0	0
2	1	3	2	0	0	0	0	0
3	4	3	1	4	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Table 4.6
Object 2 - 50 fixation learning

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	1	0	0	0	0	0	0
1	0	17	3	4	0	0	0	0
2	0	4	4	0	0	0	0	0
3	0	0	1	4	0	0	0	3
4	0	0	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	1	2	0	0	1	0	0	3

No	O^{10}
1	3 0 0 0 3 3 3 1 1 1
2	1 2 2 0 1 2 3 0 0 0
3	0 3 3 3 3 0 1 1 1 1
4	1 1 3 3 0 0 0 3 3 3
5	3 1 0 1 2 1 1 2 3 0
6	0 0 3 3 3 1 1 0 1 2
7	1 1 1 2 3 0 0 0 1 3
8	2 3 0 0 0 3 3 3 1 3
9	1 1 1 1 2 2 2 1 1 2
10	0 0 0 0 3 3 3 3 1 1
11	1 1 2 1 1 3 0 0 0 3
12	0 0 2 1 2 1 2 3 2 3
13	0 0 0 0 3 1 3 1 1 0
14	1 2 1 1 2 0 0 0 0 3
15	0 1 0 3 2 2 3 0 1 1
16	2 2 3 1 2 0 0 0 0 3
17	3 3 3 0 2 2 1 1 1 0
18	1 2 1 1 1 3 0 0 0 0
19	0 0 3 3 3 1 1 1 1 0
20	1 2 1 1 2 0 0 0 0 3

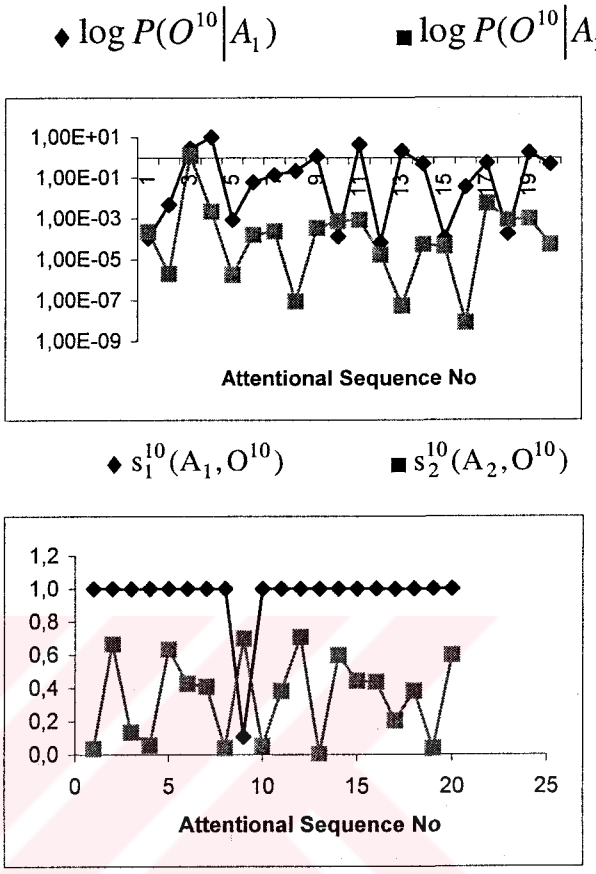

Figure 4.8 Results after 10 fixations on Scene 1 after 50 fixation learning on Scene 1 and Scene 2. Recognition rate is 85% with Markov models and 95% with evidential reasoning.

No	O^{10}
1	7021135374
2	0133311111
3	1111111177
4	7142111133
5	1161111473
6	1273311112
7	1177331111
8	1313711122
9	1477331111
10	3111121117
11	0142111511
12	1333371173
13	7702333371
14	4211111121
15	1077331111
16	1214113337
17	0123111121
18	1477335111
19	0121111174
20	1173311611

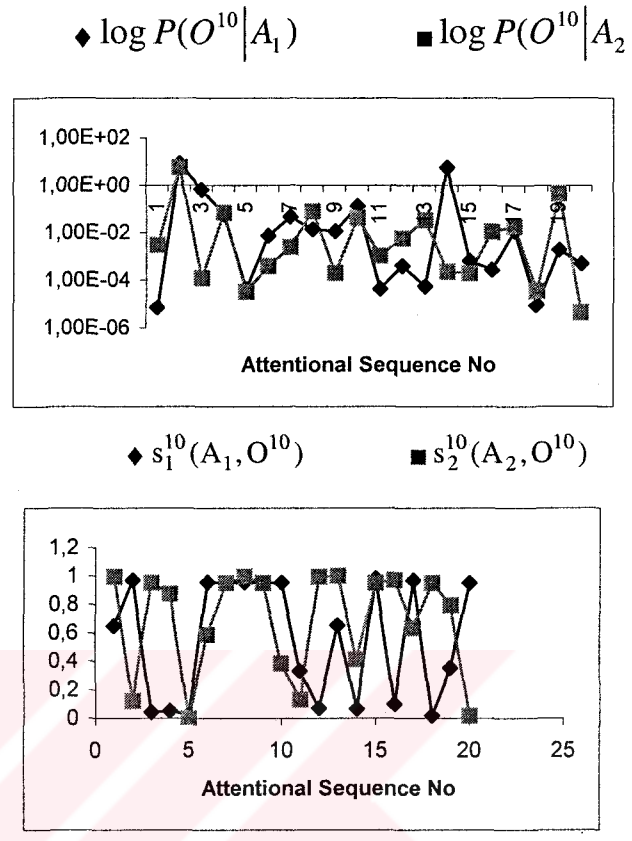

Figure 4.9 Results after 10 fixations on Scene 2 after 50 fixation learning on Scene 1 and Scene 2. Recognition rate is 50% with Markov models and 60% with evidential reasoning.

4.4.2 Complex Scenes

In the next set of experiments, 3 complex scenes shown in Figure 4.10 from our laboratory were used. Fixation points and foveas are at the center of each visual field image. Figure 4.11 and Figure 4.12 show visual fields of APES for two sample fixation sequences – looking at Scene 1. The complexity of our problem can be observed in these sample sequences. For example in the fifth fovea, a boundary caused by a shadow is fixated, and in some foveas like those numbered 4,8,9, and 10 the image is distorted by small camera or body motion, making edge based features quite hard to detect correctly. Note that these are problems common to any practical implementation outside controlled environments. Our methods are expected to cope with such distortions. Also note that in the two sequences, although starting points are close and the first visual fields are almost identical, the two sequences are quite different. However spatial and temporal relations of

observed features remain the same. One of the main contributions of our work is to develop methods for detecting these invariant relations.

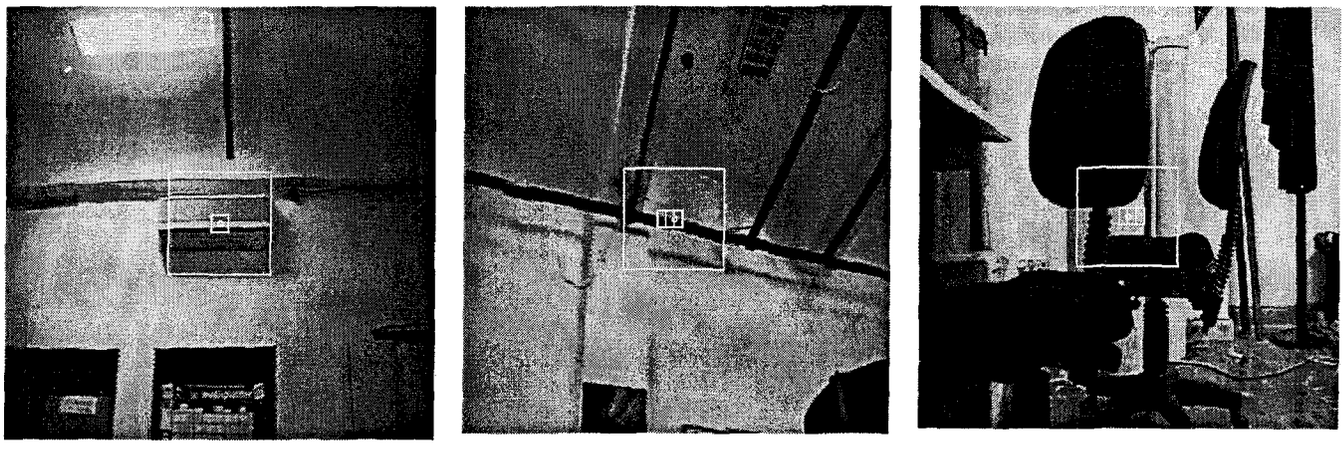

Figure 4.10 (Left to right) Wide-angle images of Scene 1, Scene 2 and Scene 3. Squares represent the visual field and fovea.

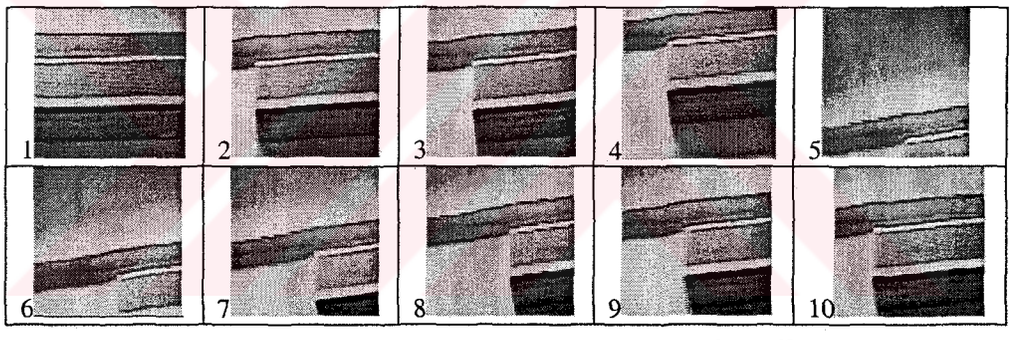

Figure 4.11 A sample sequence of visual field images $I_v=(I_v^1,....I_v^{10})$ on Scene 1.

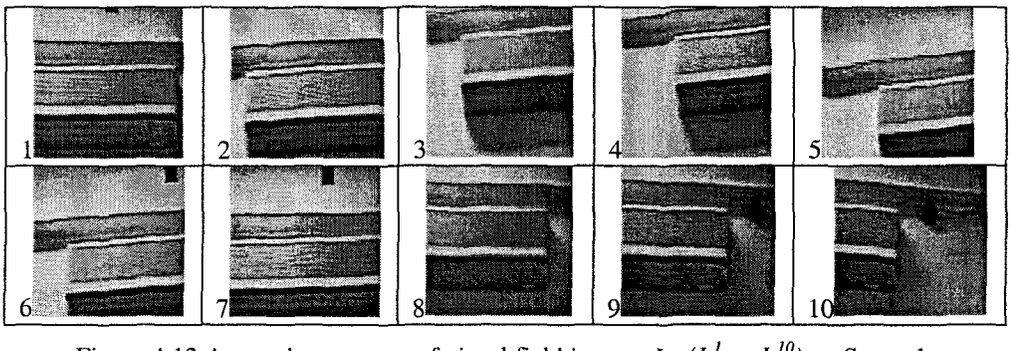

Figure 4.12 A sample sequence of visual field images $I_v=(I_v^1,....I_v^{10})$ on Scene 1.

We then compared responses using pairs of models - using these complex scenes. Their models were learned using attentional sequences of length 30. Table 4.7, Table 4.8 and Table 4.9 give the feature transition frequencies for the three scenes. Simply looking at the generated model matrices, it is observed that Scene 3 has unique features as compared

to both Scene 1 and Scene 2. Therefore any sequence generated on Scene 3 is likely to be identified correctly. On the other hand Scene 1 and Scene 2 models are very similar making classification almost impossible. These results are justified in Figure 4.13 to Figure 4.18. Scene 3 is recognized with a rate of 100% in all cases as it has a dominating model. In experiments on Scene 1 and Scene2, results of the two approaches are inconsistent.

Table 4.7
Scene 1

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	7	5	0	1	0	0	0	0
1	5	11	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Table 4.8
Scene 2

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	3	2	2	4	0	0	0	0
1	3	6	0	0	0	0	0	1
2	3	0	0	0	0	0	0	0
3	2	2	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0

Table 4.9
Scene 3

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	1	0	1	0	0	2	0
1	1	2	1	2	0	0	0	0
2	0	2	4	3	0	0	0	0
3	3	1	3	0	0	0	1	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	1	0	1	1	0	0	0	0
7	0	0	0	0	0	0	0	0

No	O^{30}
1	11100001000110011000111111100
2	10101111001100013110300001111
3	11111000110001001110110400 1110
4	00000111111111111110102011010 0
5	10001000010113111111301111 1011
6	11160110000000011200000001003 00
7	10001011101010101010000111 1013
8	00000000111113011110010111 0000
9	11101411000111000000600011111 0
10	10101010111000100000111110 0100

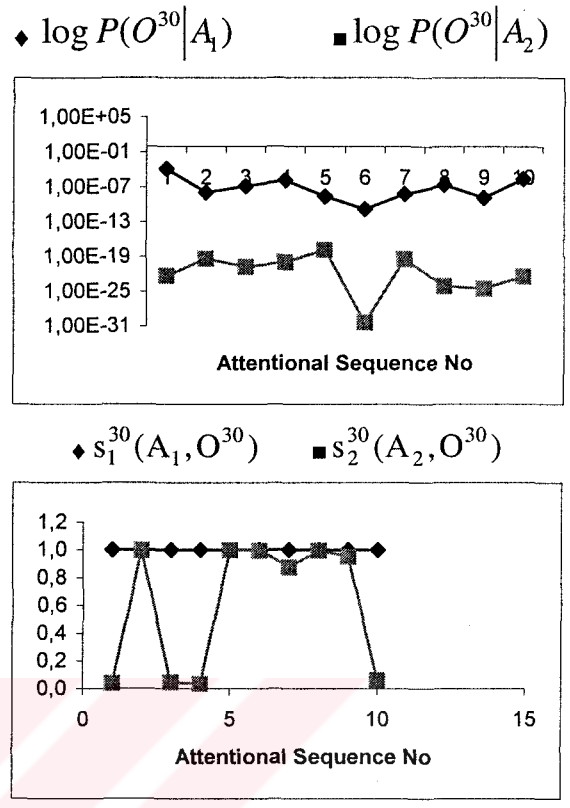

Figure 4.13 Results after 30 fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 3. Recognition rate is 100% with Markov models and 100% with evidential reasoning.

No	O^{30}
1	21201311111203123321012032102 6
2	32731323152131316035566221102 2
3	30110111100111323312101111133 1
4	00026611013131312113322321411 1
5	11112666102501131222361102222 0
6	11211313322022313223243622621 1
7	33111126116311111121311121222 2
8	11322012111101112226232222233 3
9	12330232212163301112136122221 2
10	36223123322601133011123302711 6

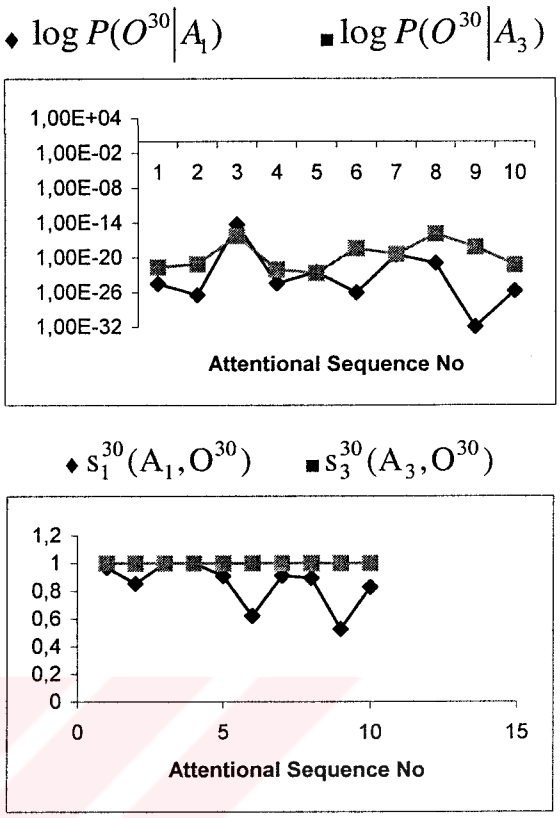

Figure 4.14 Results after 30 fixations on Scene 3 with 30 fixation learning on Scene 1 and Scene 3. Recognition rate is 80% with Markov models and 100% with evidential reasoning.

No	O^{30}
1	11100001000110011000111111100
2	10101111001100013110300011111
3	11111000110001001110110400111 10
4	00000111111111111101020111 0100
5	10001000010113111111301111 1011
6	11160110000000112000000010 0300
7	10001011101010101010000111 1013
8	00000000111113011110010111 0000
9	11101411000111000000600011 1110
10	10101010111000100000111110 0100

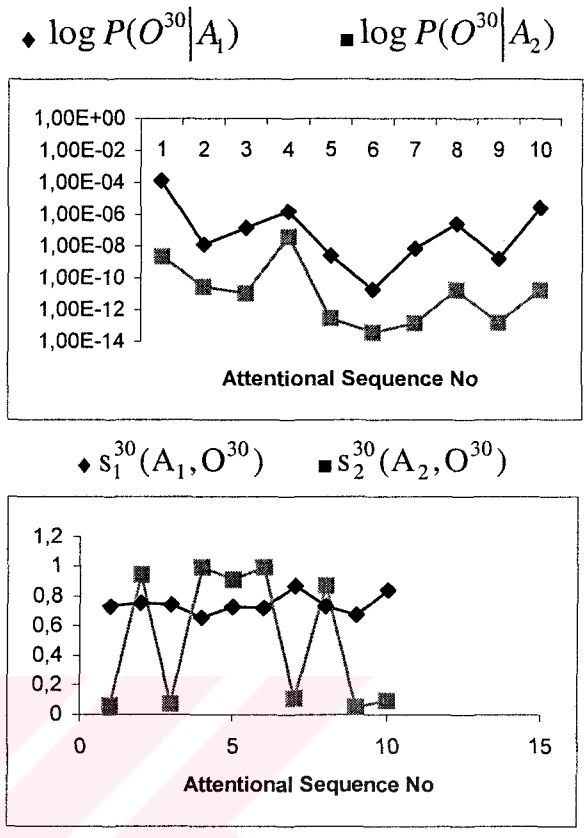

Figure 4.15 Results after 30 fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 100% with Markov models and 50% with evidential reasoning.

No	O^{30}
1	10100110706006401115441122512
2	50133330007760611515414134044 2
3	11630333111106161601067463011 0
4	01211101305170717064401110000 1
5	12230015001000176763305544132 4
6	26111701100103213200100123021 1
7	00101010371010033311011000111 1
8	00010000711101013300211110003 1
9	01110431130111100333312011001 0
10	10000111117011011110011001210 0

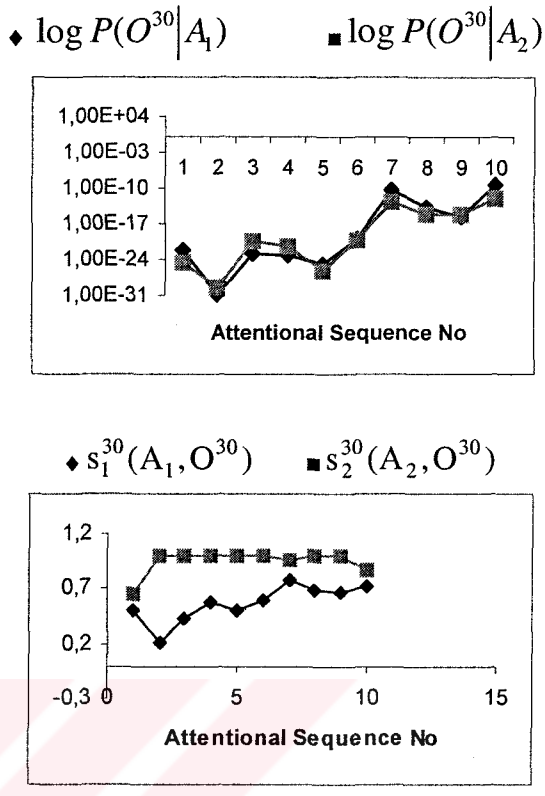

Figure 4.16 Results after 30 fixations on Scene 2 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 40% with Markov models and 100% with evidential reasoning.

No	O^{30}
1	10100110706006401115441122512
2	50133330007760611515414134044 2
3	11630333111106161601067463011 0
4	01211101305170717064401110000 1
5	12230015001000176763305544132 4
6	26111701100103213200100123021 1
7	00101010371010033311011000111 1
8	00010000711101013300211110003 1
9	01110431130111100333312011001 0
10	10000111117011011110011001210 0

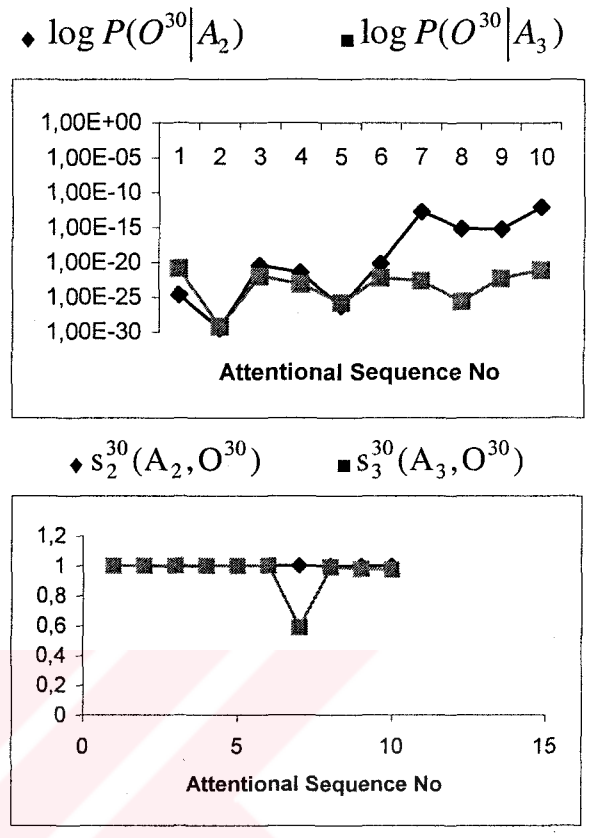

Figure 4.17 Results after 30 fixations on Scene 2 with 30 fixation learning on Scene 2 and Scene 3. Recognition rate is 70% with Markov models and 70% with evidential reasoning.

No	O^{30}
1	21201311111203123321012032102 6
2	32731323152131316035566221102 2
3	30110111100111323312101111133 1
4	00026611013131312113322321411 1
5	11112666102501131222361102222 0
6	11211313322022313223243622621 1
7	33111126116311111121311121222 2
8	11322012111101112226232222233 3
9	12330232212163301112136122221 2
10	36223123322601133011123302711 6

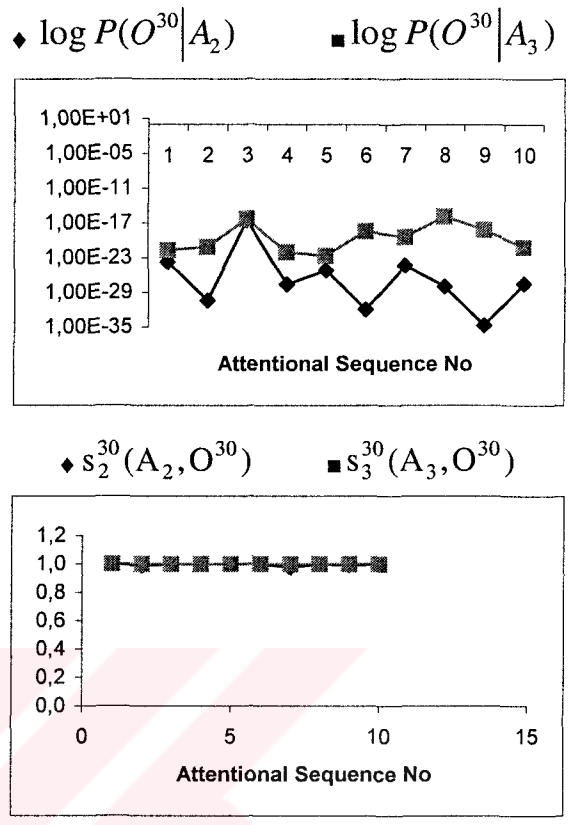

Figure 4.18 Results after 30 fixations on Scene 3 with 30 fixation learning on Scene 2 and Scene 3. Recognition rate is 100% with Markov models and 100% with evidential reasoning.

4.4.3 Complex and Similar Scenes

The method of evidential reasoning was also tested using three similar scenes with small variations and one unrelated scene. Changes in the scene are not very small at all, such as missing chairs, but a human viewer tends to overlook these changes. APES is expected to perform similarly and "understand" that the three scenes belong to the same part of the world and the fourth scene to a different part. The four scenes are shown in Figure 4.19.

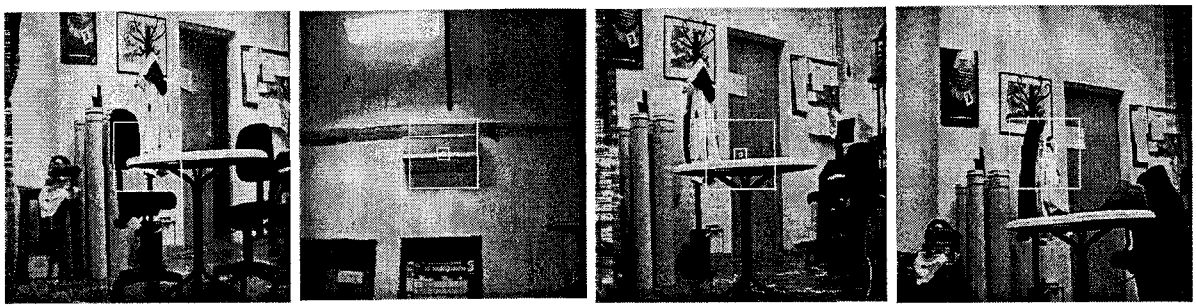

Figure 4.19 (Left- to right) Wide-angle images of Scene 1, Scene 2, Scene 3 and Scene 4.

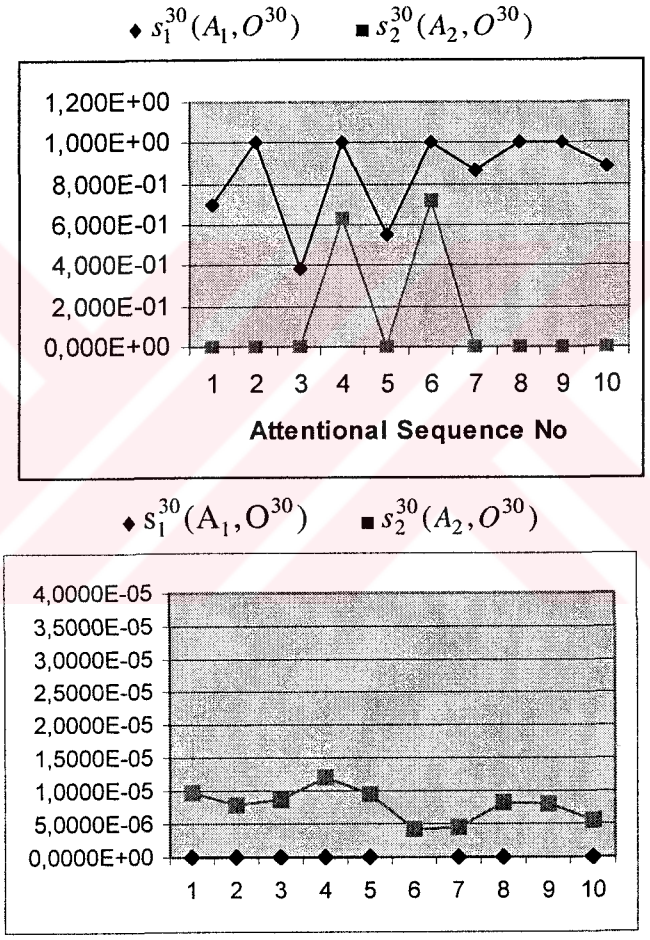

Figure 4.20 Results of 30 fixations on Scene 1 (top) and Scene 2 (bottom) after 30 fixation learning on Scene 1 and Scene 2. Recognition rates are 100% and 80% respectively.

In Figure 4.20 results of experiments on the original training scenes are shown. Scene 1 can be recognized easily with a high margin, while Scene 2 is recognized in 80% of the experiments with a very low margin. In Figure 4.21 results of experiments on the two variants of Scene 1, Scene 3 and Scene 4 are shown. Both scenes can easily be recognized as Scene 1 except in a few experiments.

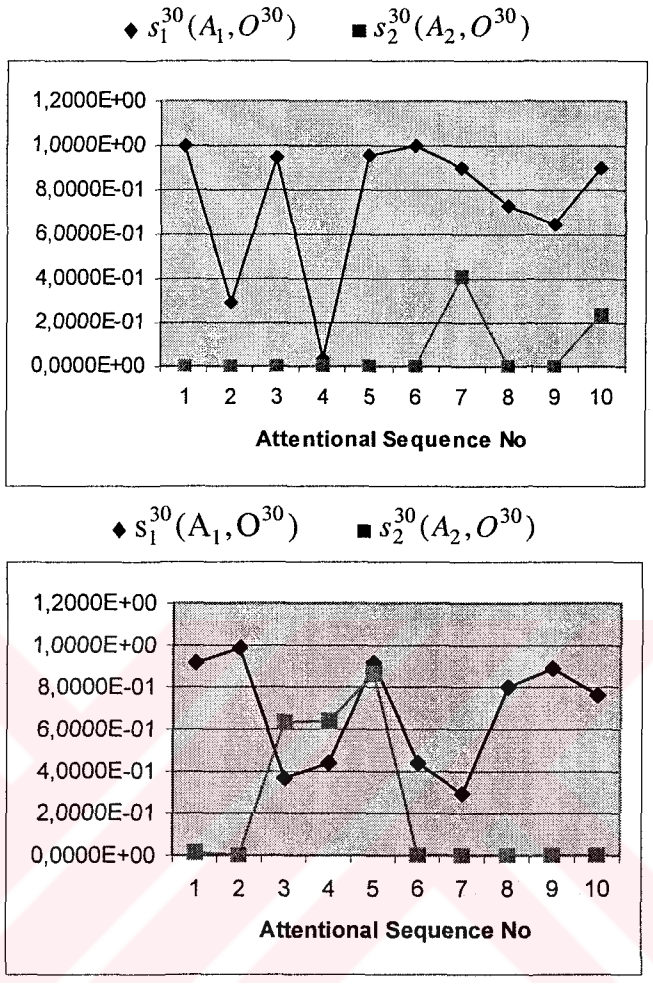

Figure 4.21 Results of 30 fixations on Scene 3 (top) and Scene 4 (bottom) after 30 fixation learning on Scene 1 and Scene 2. Recognition rates are 100% and 80% respectively.

Although these experiments show that scene recognition based on attentional sequences can compensate for small changes in the environment, the low margins in Scene 2 recognition results in Figure 4.20 is confusing. This result may suggest that the model of Scene 1 may be dominating over Scene 2 and correct classification of Scene 3 and Scene 4 is a result of this dominance.

4.4.4 Results and Comparison

In summary, our experiments on simple and complex scenes revealed the following important results about the use attentional sequences for scene classification:

1- Both Markov models and evidential reasoning are promising for classification of attentional sequences.

2- Even by using very simple edge based features we can deduce invariant relations from the seemingly varying fovea image sequences generated while looking at the same scene.

3- Using as low as 10 fixations during learning and recognition, good classification performance can be achieved using both methods.

4- Results on complex real world scenes, which are hard to classify using classical methods, show that attentional sequence based classification is promising to solve such problems.

5- Increasing the learning period does not necessarily improve performance. Good performance with short learning period is possible depending on learning and recognition fixations.

6- The two models performed similarly in simpler classification tasks, where models were distinct. In harder tasks either both methods generated very small margins between the two models and returned false results, or evidential reasoning performed better. The differences between the two methods are caused by the fact that unlike Markov methods contributions from competing models are taken into account by the combination rules used in evidential reasoning.

7- In order to achieve good performance, models (feature transition frequency matrices) need to represent unique features about the scene. How to generate fixation models with such property and how to compute their representation capability are open problems we are working on.

4.4.5 Discussion

The main objective of our work was to investigate whether the attentional sequence can be used for scene classification by applying the above methods. Therefore, in order to reduce the effects of attention mechanism, simple attentive features are used in our experiments. However, the behavior of the system can be controlled effectively by using different attentional schemes including top down approaches, although how this should be done is an open question. In general the performance of sequence classification will be

unaffected as long as the same deterministic attention scheme is used during both modeling and recognition. However, stochastic components in the attentional scheme may change the performance as the classification algorithms rely on the observation of learned sequences or short segments of learned sequences.

The use of only 8 simple edge features in our experiments is also restrictive. As seen in the experiments different scenes may lead to similar models, which do not have any discriminating ability. Instead, using many complex features in the attentional sequence and the spatial locations of features can improve performance. Especially in complex scene experiments a better model of the environment can be obtained. However the detailed scene models generated in this way may also be restrictive and the generalization behavior demonstrated in the experiments of section 4.4.3 may not be achieved.

As mentioned above, one of the main strengths of our approach is the ability to change pre-attentive and attentive features as well as the attention scheme without changing the sequence modeling and classification methods. Therefore an adaptive system can modify these sub-systems based on the current task specification while keeping the same decision system.

5. BUBBLE MODEL

According to the widely accepted *integrative visual buffer theory* of Breitmeyer, visual memory is a buffer which is capable of integrating visual information obtained from different spatial locations at different times by attentive processing [80,81,82]. Also the fact that we can easily recall what we have seen, where and when, proves that a spatial memory mechanism is very much involved in vision. Furthermore, in spite of continuous eye, head and body motions, humans can perceive a stable image of their environment. The mechanisms of visual integration, which are thought to be responsible from this illusion, are one of the most challenging problems in cognitive psychology [4,34,80,81,82].

5.1 Requirements for Visual Memory

Consider an attentive vision system which is free to move in 3D space. Its perceptual mechanism should integrate information in the course of active vision behavior. Suppose that the camera is fixating a location A and something in location B attracts its attention. Then, a saccade to bring the fovea on B and leaves A in the left peripheral field. If a stable, orbital centered system of reference is available, then the robot should know that what is now in sharp focus on the fovea is in the same spatial location occupied by the stimulus that has occasioned the saccade. Conversely, what was imaged on the fovea in A has now become a somewhat blurred image in the left field, but has not changed its spatial position. These relations hold only for that particular viewpoint occupied by the system.

The visual memory model should allow effective representation and use of these and similar kinds of information required by an attentive system to build a functional environment map. The information represented by this model should not be the 2D retinal image, but rather it should contain foveal processing results. Finally, depending on the visual task, the model should be able to recall different properties of the scene.

5.2 Bubble Model of Integrative Visual Buffer

Based on the integrative visual buffer theory and supporting biological evidence, we introduce the bubble model as a new way of representing, storing and using information collected during selective perception process. The bubble model enables the construction of a functional map represented in a spherical coordinate system attached to the viewpoint. Mathematically the bubble is a deformable 3D surface, which is controlled by a number of points that are coincident with the potential fixation points of the visual system. The information represented by the bubble is not the 2D retinal image but rather it comes from computations on a neighborhood of fixation points. During visual exploration the bubble is inflated via its control points depending on the information visually obtained from the corresponding fixation area. In general a set of bubbles can be used to store different types of information that can be extracted by the vision system. Bubbles are attached to the viewpoint and a new set of bubbles is used for different viewpoints. When the system visits a previously visited viewpoint the corresponding bubble set can be recalled. Although all previous visual experience is stored in a long-term memory, only functional information for the current viewpoint is recalled as required by the current task.

5.3 Potential Fixation Points and Bubble Points

Consider a camera positioned on a two degree of freedom pan-tilt base. Assuming no practical restrictions, the optical axis of the camera can be theoretically directed in any direction (θ, φ) where θ, φ are pan and tilt angles respectively. If we assign a quantitative measure $\rho \varepsilon R$ to each fixation direction, a surface is defined implicitly by the set of (ρ, θ, φ). For a constant ρ, this surface is spherical in nature. We refer to this spherical surface - hypothetically placed around the robot - as *the bubble*. Motivated by the fact that in human beings, two fixations are almost never adjacent due to the high foveal acuity, the bubble need not to be a continuous surface. Rather it can be discreet and we can use a limited number of equally spaced points on the sphere which we call *bubble points*. By directing its camera to these points the system can fixate on the corresponding points in 3D space, which we call *potential fixation points*. While the term bubble point represents an

element of our abstract representation, a potential fixation point is a real point in 3D space. In Figure 5.1, two bubble points and the corresponding potential fixation points are shown.

5.4 The Bubble

The bubble points are determined by the saccade precision in pan and tilt directions. All bubble points (θ,φ) can be defined in spherical coordinates by,

$$\begin{aligned}\theta &= 0,\Delta\theta,2\cdot\Delta\theta,3\cdot\Delta\theta,\ldots,(n-1)\cdot\Delta\theta \\ \varphi &= 0,\Delta\varphi,2\cdot\Delta\varphi,3\cdot\Delta\varphi,\ldots,(m-1)\cdot\Delta\varphi\end{aligned} \tag{5.1}$$

where $\Delta\theta$, $\Delta\varphi$ are the fixation resolutions and n, m are the number of bubble points in pan and tilt directions. Therefore the set of points that form the bubble can be defined as follows,

$$\begin{aligned}&\beta = \{(\rho,\theta,\varphi)\in \Re^3 \mid \theta \equiv i\cdot\Delta\theta, \varphi \equiv j\cdot\Delta\varphi\} \\ &where\ i\in[0,n), j\in[0,m)\end{aligned} \tag{5.2}$$

Note that we can store each bubble in a 2D array where each element corresponds to a bubble point. Then we have a *bubble matrix*

$$B = \begin{bmatrix} B_{11} & B_{12} & .. & .. & B_{1m} \\ B_{21} & B_{22} & .. & .. & .. \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ B_{n1} & B_{n2} & .. & .. & B_{nm} \end{bmatrix} \tag{5.3}$$

such that,

$$\forall((\rho,\theta,\varphi)\in\beta) \Leftrightarrow \exists(B_{ij}\in B \mid \theta \equiv i\cdot\Delta\theta, \varphi \equiv j\cdot\Delta\varphi) \tag{5.4}$$

We will explain the elements of the bubble matrix in the next section in greater detail, but at this stage all we need to know is that there is a one to one correspondence between potential fixation points, bubble points and the elements of the bubble matrix.

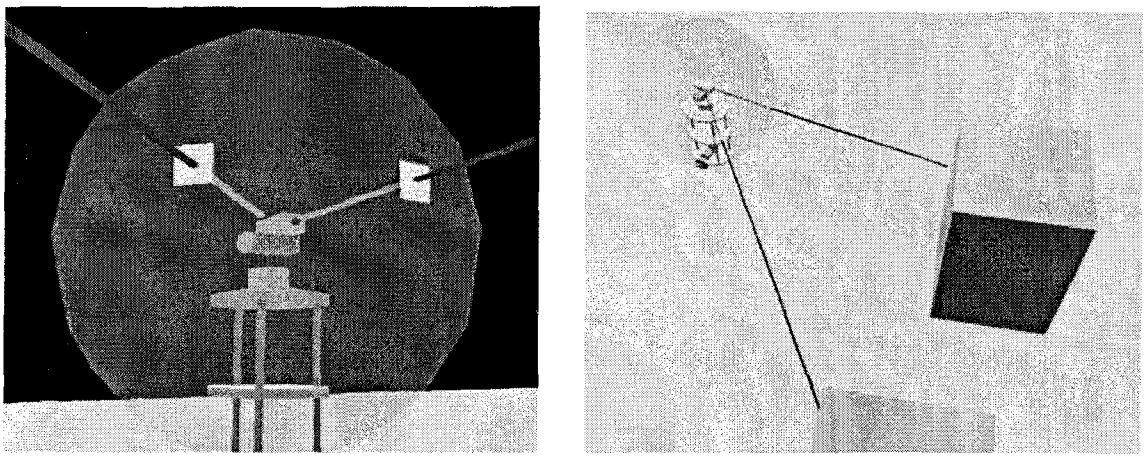

Figure 5.1 Bubble points and potential fixation points.

5.5 Bubble Functions

Suppose now that we deform the bubble at bubble points – where the amount of deformation is determined by the quantitative measure obtained from the high resolution processing made on the fovea. This area is defined in 2D image coordinates as an FxF pixel region at the centre of the acquired image. Since each fixation point is associated with a unique fovea, we can think of visual information as being concentrated at the potential fixation points. For the moment this enables us to forget about the 2D image and think that each time a fixation is made visual information representing the corresponding area is somehow extracted and assigned to the fixation point. The result is an implicit 2D surface deformed to represent information obtained from the current viewpoint. We say that "visual information inflates the bubble" and the resulting surface can be represented as:

$$\beta = \{(\rho(\theta,\varphi),\theta,\varphi) \in \Re^3 \mid \theta \equiv i \cdot \Delta\theta, \varphi \equiv j \cdot \Delta\varphi \}$$
$$where\ i \in [0,n), j \in [0,m) \tag{5.5}$$

The function $\rho:(\theta,\varphi)\rightarrow R$ is referred to as the *bubble function*. The deformation concept is illustrated in Figure 5.2 on a circle.

For example, for a simple linear measure at each fixation point (θ,φ), $\rho(\theta,\varphi)$ can be computed from the image by running an operator over the fixation fovea as follows,

$$\rho(\theta,\varphi)=\sum_{x=-F/2}^{+F/2}\sum_{y=-F/2}^{+F/2}\sum_{k=-N/2}^{+N/2}\sum_{l=-M/2}^{+M/2}I(x-k,y-l)\cdot T(k,l) \tag{5.6}$$

where T is an NxM feature template, F is the fovea size and I(x,y) is the image obtained when the camera is fixated on the potential fixation point $[\theta \quad \varphi]^T$. Note that in the above equation we are mapping the visual information contained in a fovea sized FxF region of the image to a single bubble point.

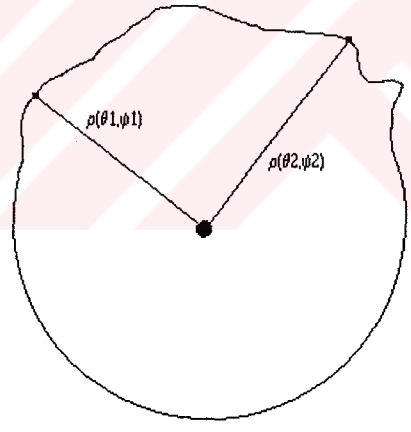

Figure 5.2 Deformed bubble representing the bubble function ρ for a single visual feature.

Note that this contraction mechanism is a property of both robotic and biological active vision systems. The bubble function $\rho(\theta,\varphi)$ can be used to represent any visual feature or a combination of visual features. Furthermore, multiple bubbles can be formed - each corresponding to one distinct combination of features. Then the set of all bubbles which can be computed using available resources provide a compact representation of all the information that the system can visually extract from its environment while standing at a specified viewpoint.

Mapping visual information from image to bubble is an interesting problem, which deserves some more attention. In our approach the radii of bubble control points represent visual information and exactly one ρ value corresponds to each fixation point. These values can be organized in the form of a 2D array, which is the bubble matrix we defined above. The elements of the B matrix can simply be computed by,

$$B_{ij} = k \cdot \rho(i \cdot \Delta\theta, j \cdot \Delta\varphi) \tag{5.7}$$

where k is a scaling constant used to scale the visual feature represented by the bubble function. Since there may be many bubbles for one visual space each representing a different type of information or modality the constant k can be used to normalize bubble functions. More complex strategies using features at different resolutions and abstraction levels can also be applied to map visual information from the image to the bubble depending on what needs to be stored.

5.6 Contributions of the Model

The bubble model is proposed as a convenient way of storing and using information collected by active vision systems. The benefits of using this representation are listed below:

- Human vision is based on selective attention, which requires integration of spatially distinct features in time to obtain a model of the environment. The bubble approach can be an acceptable model for this integration behaviour.
- Visual information is originally obtained in 2D cartesian coordinates which is not the natural coordinate system of an active vision system. The bubble representation enables us to manipulate visual information in spherical camera coordinates
- In a moving vision system visual information is heavily dependent on the viewpoint. Bubble centered coordinates enable us to distinguish and store information obtained from different viewpoints. This can also be used to extract depth with a single sensor.

- Different visual features can be stored in separate bubbles. If this information can be managed intelligently, conflicts during high level tasks can be solved.
- During selective attention using bubbles to store saliencies of candidate foveas in the periphery can reduce peripheral computations up to 50% if the same area falls into the periphery at a later fixation.
- Bubbles can have a time stamp which enables comparison or updating based on viewing times. This makes the representation suitable for dynamic visual environments.

5.7 Implementation

The bubble model is implemented on APES. A number of experiments to test the bubble formation in different situations are made. In this case the task, borrowed from cognitive psychology literature, is defined as visual feature integration over saccades, where the ability to integrate features obtained from different spatial locations at different times is tested. By definition, this task is easily accomplished using a bubble memory. For the scene shown in Figure 5.3, the deformed bubble is shown in Figure 5.4. In this experiment the part of the scene containing the bars is an interesting area for our system which uses an edge based attentive criteria. Saccades are vertical in nature and multiple fixations are made on this part of the scene. We also observed significant changes in the strength of features computed from the same point in this scene, which was due to the changing lighting conditions outside the window.

Figure 5.3 Experiment 1- Window-with-bars scene.

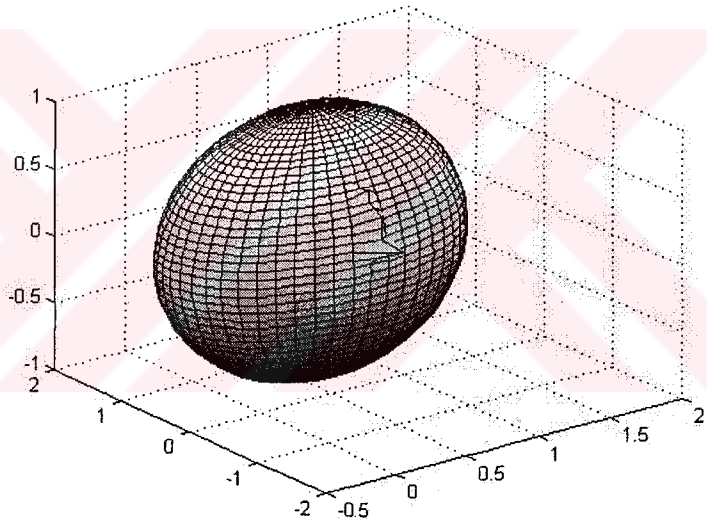

Figure 5.4 Bubble trace for fixations on Window-with-bars scene.

In the next experiment a crowded scene from our laboratory is used as shown in Figure 5.5. In this case, there is a greater number of points worth attending to. In the two bubble traces shown in Figure 5.6, fixations are concentrated on the robot in the with some fixations also to the left and right. As the exploration continues and all possible fixation points are visited, the bubbles are expected to converge to a single representation of the environment from this viewpoint regardless of the starting point and sequence of fixations being made.

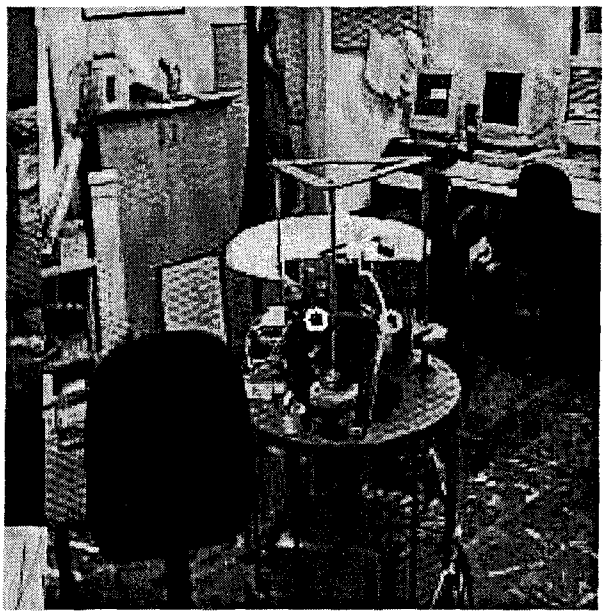

Figure 5.5 Laboratory scene.

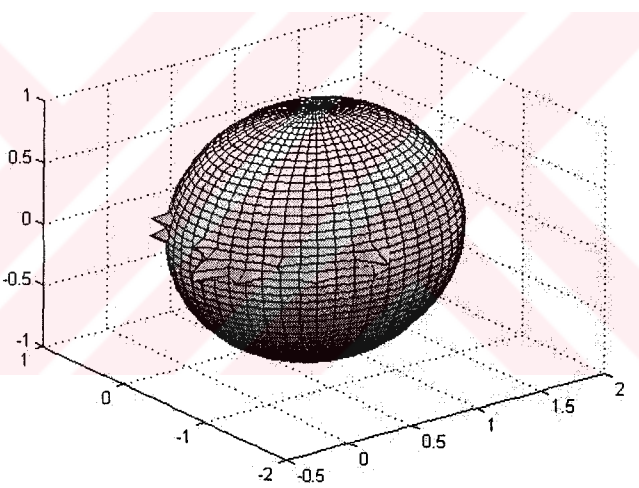

Figure 5.6 Bubble trace for fixations on Laboratory scene.

5.8 Scene Recognition with Bubbles

Next we investigate the possibility of using bubbles for vision based modeling and recognition of 3D environments. The integration property of bubbles enable them to store foveal features observed from a single point in space. This representation can be used to model and recognize 3D environments in a short time if the bubble forming sequence of fixations are similar.

In the next set of experiments we use the library scene from our laboratory shown in Figure 5.7. In Figure 5.8 the gradient of library scene which gives an idea about edge based saliency computations is shown. Throughout these experiments the first fixation frame shown in Figure 5.9 is the same, enabling us to observe the differences in attentional sequences and their effects on bubble shapes.

Figure 5.7 Library scene from our laboratory.

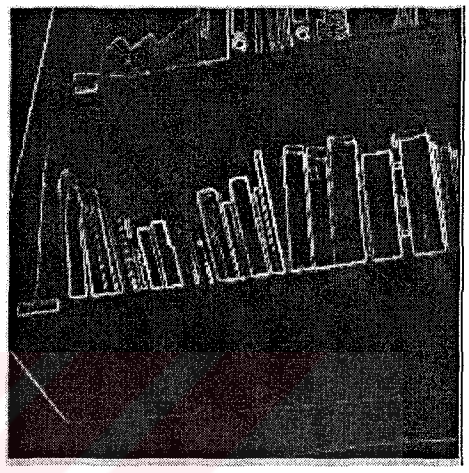

Figure 5.8 Gradient of library image.

Figure 5.9 First fixation frame used in library scene experiments.

Using saliency computations in the fovea, bubbles from 50 fixations on the library scene are formed in 6 different runs. The resulting bubbles are shown in Figure 5.10. The variations in attentional sequences are seen on bubbles, but these variations have little effect on the general bubble shape. However bubbles shown in Figure 5.12 formed while looking at a different scene in Figure 5.11 have a completely different shape. Therefore we conclude that a bubble modeling algorithm preserving this general bubble shape can be

used for environment recognition based on this visual representation. We look at this problem in section 5.9 where we use Fourier methods for storing bubble data in a compact mathematical form and at different levels of detail.

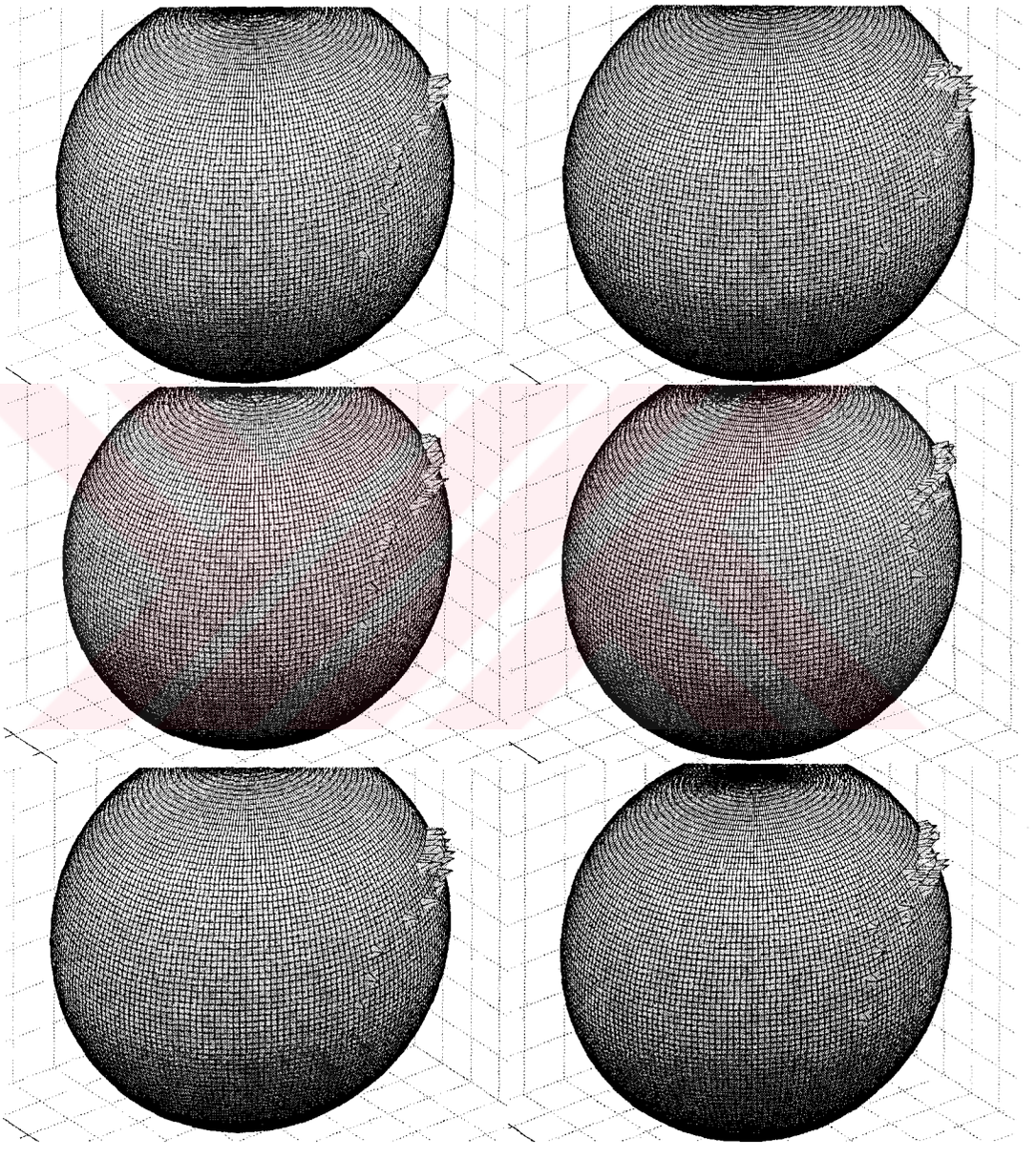

Figure 5.10 Bubbles formed in 6 different experiments on library scene.

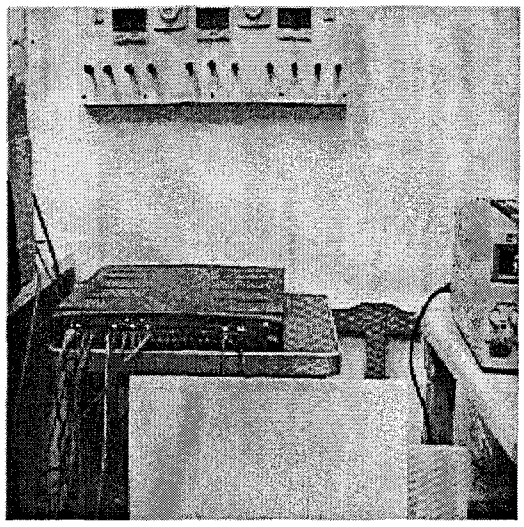

Figure 5.11 Scene containing hub and old switchboard from our laboratory.

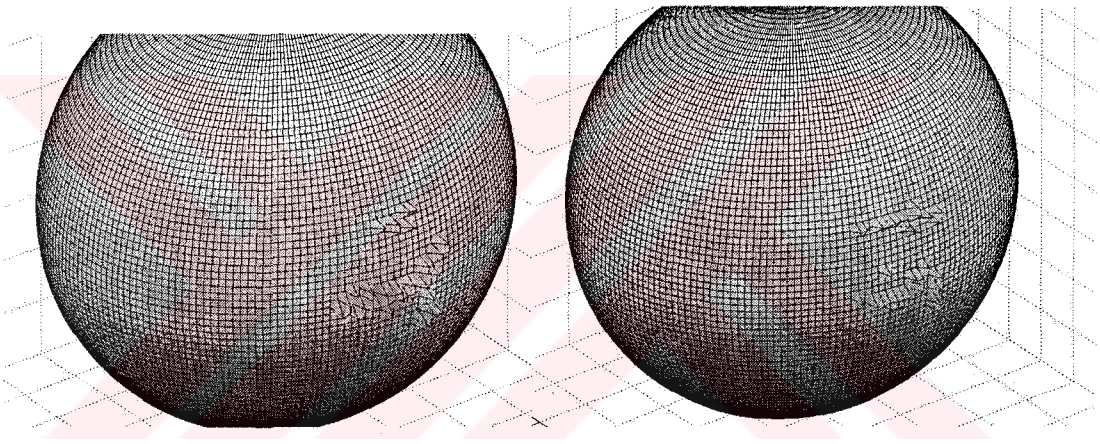

Figure 5.12 Bubbles generated in experiments on hub-switchboard scene.

In order to understand the discrimination capability of bubbles in a scene recognition task, we also design an experiment where both saliency values and evidence based supports for two different models are used to form bubble surfaces. Experiments are performed on the scene shown in Figure 5.11 containing hub and network cables, and the old switchboard mounted on the wall. A series of fixations and the corresponding bubble inflated by saliency values are shown in Figure 5.13. In this figure the horizontal and vertical fixations in the lower part are on the hub and its extending network cables, and the upper isolated group of fixations is on the old switchboard.

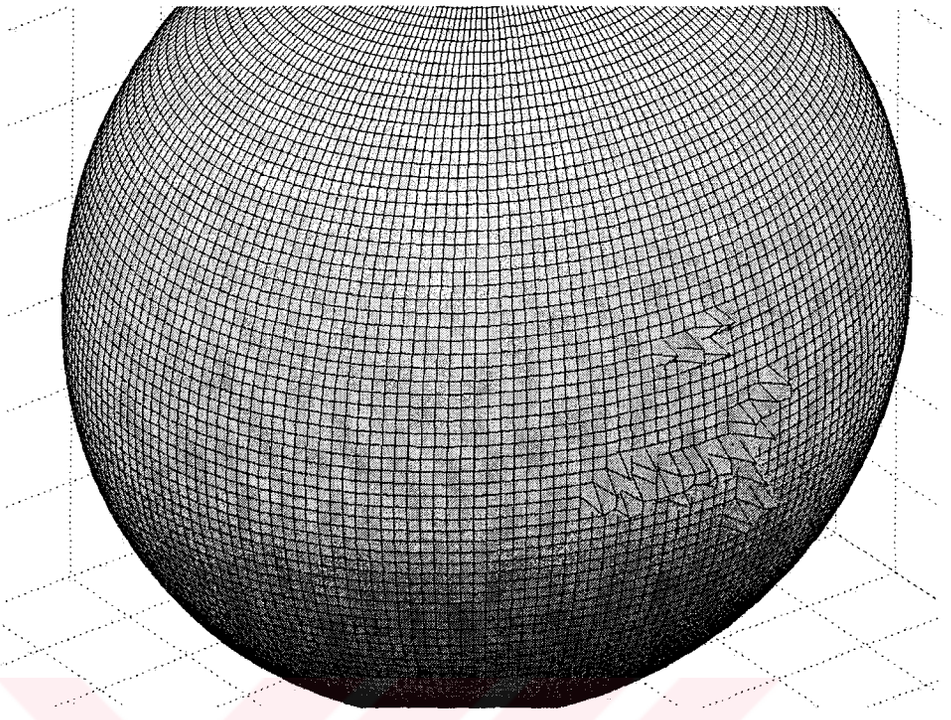

Figure 5.13 Saliency bubble.

Then, using fixations on the hub and switchboard we generate two library models, model 1 and model 2 respectively. These models and the original set of fixations are used in a recognition experiment to generate support values for the two models. Figure 5.14 and Figure 5.15 show the bubbles inflated using these support values for the two models. Using this representation we can exactly observe the amount of support provided at each stage for each model. As seen in these figures the bubble for model 1 supports has no activity around the old switchboard, and the model 2 bubble has little activity around the hub region. Looking at the two bubbles it is also seen that their shapes roughly add up to give the saliency bubble.

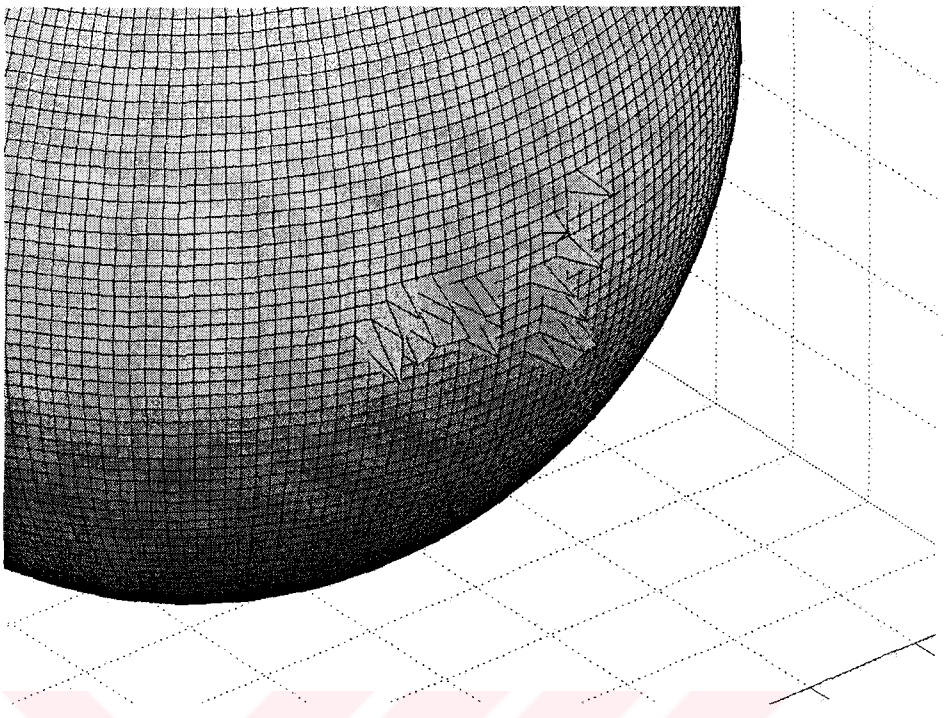

Figure 5.14 Bubble formed using supports for model 1 (hub) at each fixation.

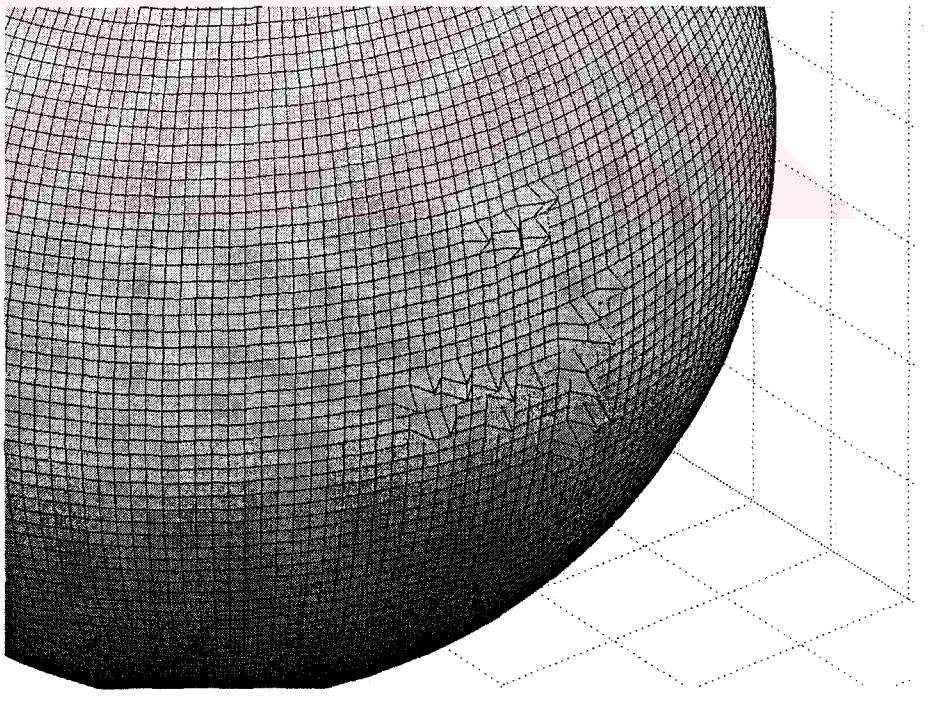

Figure 5.15 Bubble formed using supports for model 2 (old switchboard) at each fixation.

In order to confirm the validity of support bubbles we also plotted support values for the two models. In Figure 5.16 only fixations after 10^{th} are considered so that enough information will be accumulated before starting recognition decisions. Initially after 10^{th}

fixation (shown as 1 in the figure) model 1 is dominating. Starting with 20th fixation the robot starts looking at model 2 and after a transition period, where no decision is possible, model 2 is activated and model 1 goes down starting with 35th fixation. As the robot attends the two areas of the scene, the values change to support corresponding models.

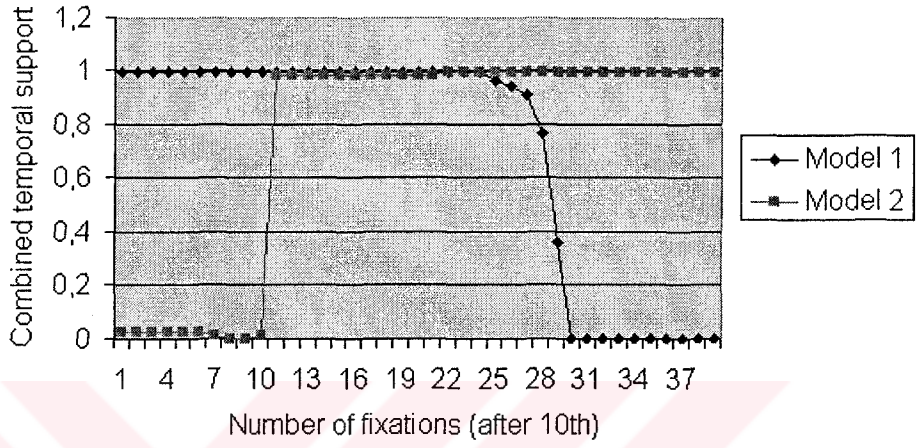

Figure 5.16 Supports for model 1 and model 2 vs. fixation number (starting from 10th)

5.9 Bubble Modeling and Reconstruction

Although using bubbles for visual environment representation is much more effective compared to working with images, still the representation may take up a lot of memory especially when a number of different bubbles are used to store different features for a given task. However, being well defined 3D closed curves bubbles can be parametrically modeled in order to simplify storage and control. This promises rapid comparison of bubbles of two nearby viewpoints or of two different scenes. Bubble models can potentially lead to an optimal and dynamic representation of the 3D visual environment.

The method we applied for bubble modeling and reconstruction is based on 3D Fourier surfaces. Fourier representations express functions in terms of an orthonormal basis. In general if $\phi_k(t)$ is such a basis then the function f(t) can be represented on the interval (a,b) by

$$f(t)=\sum_{k=1}^{\infty} p_k \phi_k(t) \quad , \qquad p_k = \int_a^b f(t)\phi(k)dt \tag{5.8}$$

and the parameters p_k are the parameters of the representation.

In the case of a single variable periodic function, sine and cosine functions form a suitable orthonormal basis. However a 3D surface is explicitly described by three coordinate functions of two surface parameters,

$$f(u,v) = (x(u,v), y(u,v), z(u,v)) \tag{5.9}$$

In order to represent these two variable coordinate functions we need to use a different basis. The following basis is successfully used in the literature:

$$\varphi = \begin{Bmatrix} 1, \cos mu, \sin mu, \cos lv, \sin lv, \cos mu \cdot \cos lv, \sin mu \cdot \cos lv, \\ \cos mu \cdot \sin lv, \sin mu \cdot \sin lv, \ \ldots \ (m,l = 1,2,\ldots) \end{Bmatrix} \tag{5.10}$$

Then a function of two variables $f(u,v)$ can be represented by

$$f(u,v) = \sum_{m=0}^{K_1-1} \sum_{l=0}^{K_2-1} \lambda_{m,l} \begin{bmatrix} a_{m,l} \cdot \cos mu \cdot \cos lv + b_{m,l} \cdot \sin mu \cdot \cos lv + \\ c_{m,l} \cdot \cos mu \cdot \sin lv + d_{m,l} \cdot \sin mu \cdot \sin lv \end{bmatrix} \tag{5.11}$$

where

$$\lambda_{m,l} = \begin{cases} 1 & for \quad m = 0, l = 0 \\ 2 & for \quad m > 0, l = 0 \ or \ m = 0, l > 0 \\ 4 & for \quad m > 0, l > 0 \end{cases}$$

and the series is truncated after K_1 -1 and K_2 -1 terms. In this representation three sets of parameters corresponding to x(u,v), y(u,v), z(u,v) are,

$$\begin{aligned} p_x &= \{a_x, b_x, c_x, d_x\} \\ p_y &= \{a_y, b_y, c_y, d_y\} \\ p_z &= \{a_z, b_z, c_z, d_z\} \end{aligned} \tag{5.12}$$

These are collectively referred to as the parameter vectors

$$p = \{p_x, p_y, p_z\} \tag{5.13}$$

The bubble can be parametrically represented by using the bubble function $\rho(\theta,\varphi)$, where the two surface parameters θ,φ are the pan and tilt angles.

When the whole basis shown above is used for surface representation a torus can be formed. To represent other surfaces a subset of the basis should be used. In the case of a closed surface the basis functions are,

$$\varphi_{closed} = \{1, \sin l\varphi, ..., \cos m\theta \cdot \sin l\varphi, \sin m\theta \cdot \sin l\varphi, ... (m,l = 1,2,...)\} \tag{5.14}$$

Using this basis a tube whose ends close up to a point is formed. However the ends are also forced together, therefore a weighted term of the form

$$\sin\left(\varphi - \pi/2\right)$$

should be added to each coordinate.

In order to test this approach modeling and reconstruction experiments are performed on a randomly inflated bubble shown in Figure 5.17.

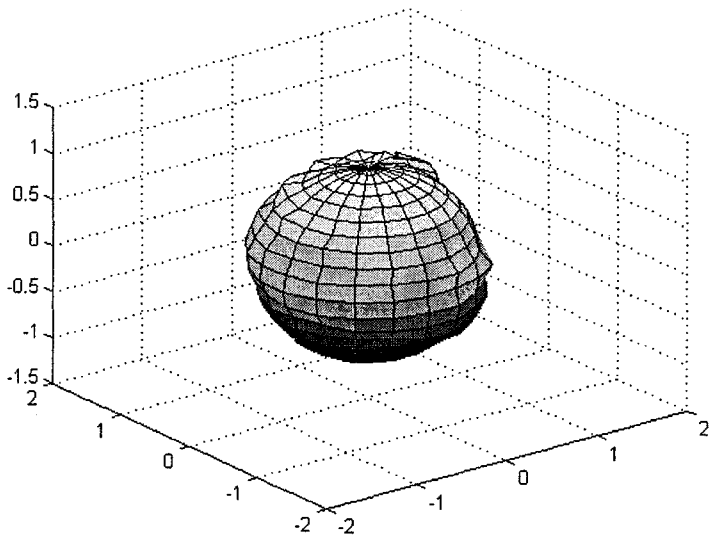

Figure 5.17 Original inflated bubble

Then reconstruction experiments are made using the above model and different number of terms or coefficients. Results of reconstruction are shown in Figure 5.18 through Figure 5.21

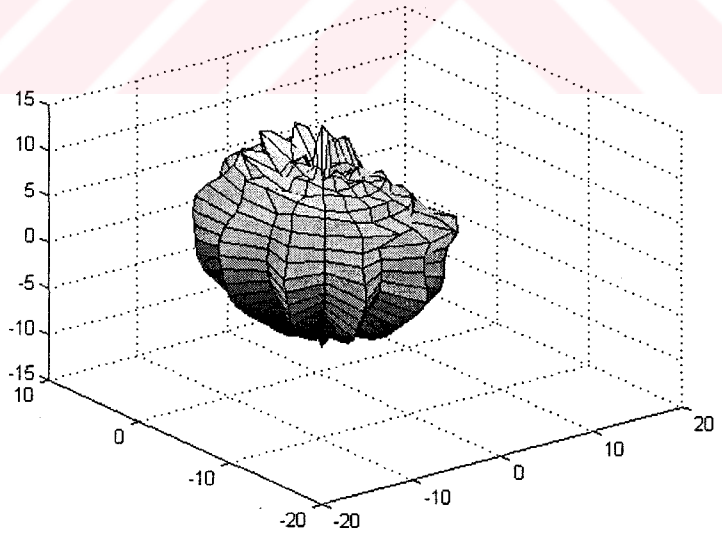

Figure 5.18 Reconstruction with 5 coefficients

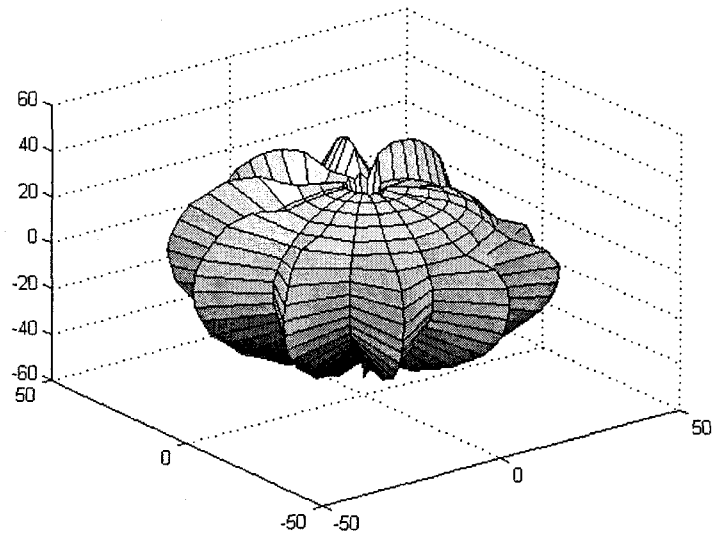

Figure 5.19 Reconstruction with 10 coefficients

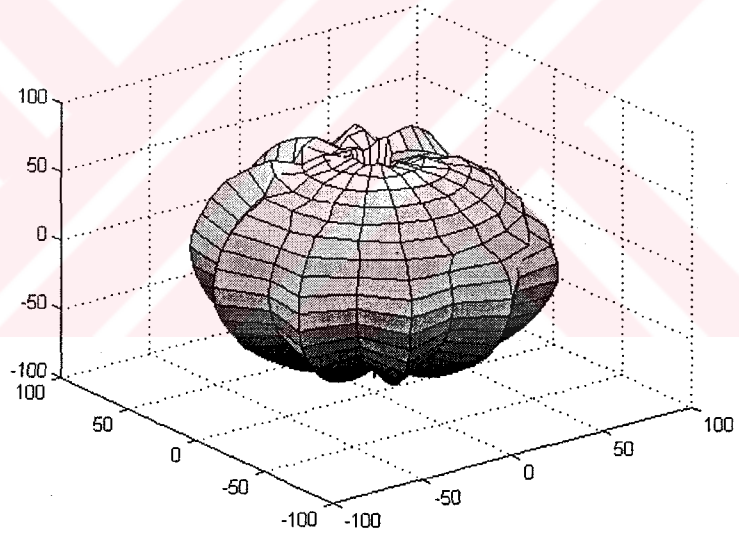

Figure 5.20 Reconstruction with 15 coefficients

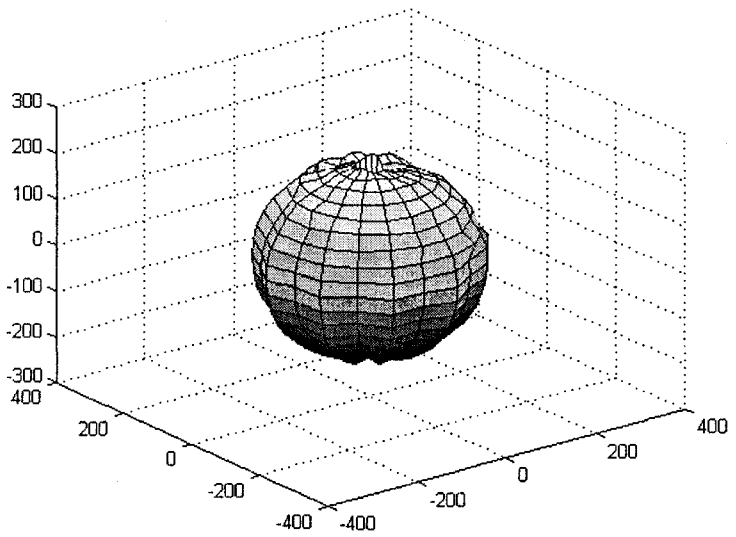

Figure 5.21 Reconstruction with 25 coefficients

6. AN INTEGRATED MODEL OF ATTENTIVE VISION

The attention model, sequence processing methods, and memory mechanisms developed above can be combined into an integrated model of active and attentive vision. This model simulates some important properties of human vision, like fovea-periphery distinction, selective attention, and serial processing. It also proposes models for other poorly understood mechanisms, like temporal recognition, visual integration, and environment modeling.

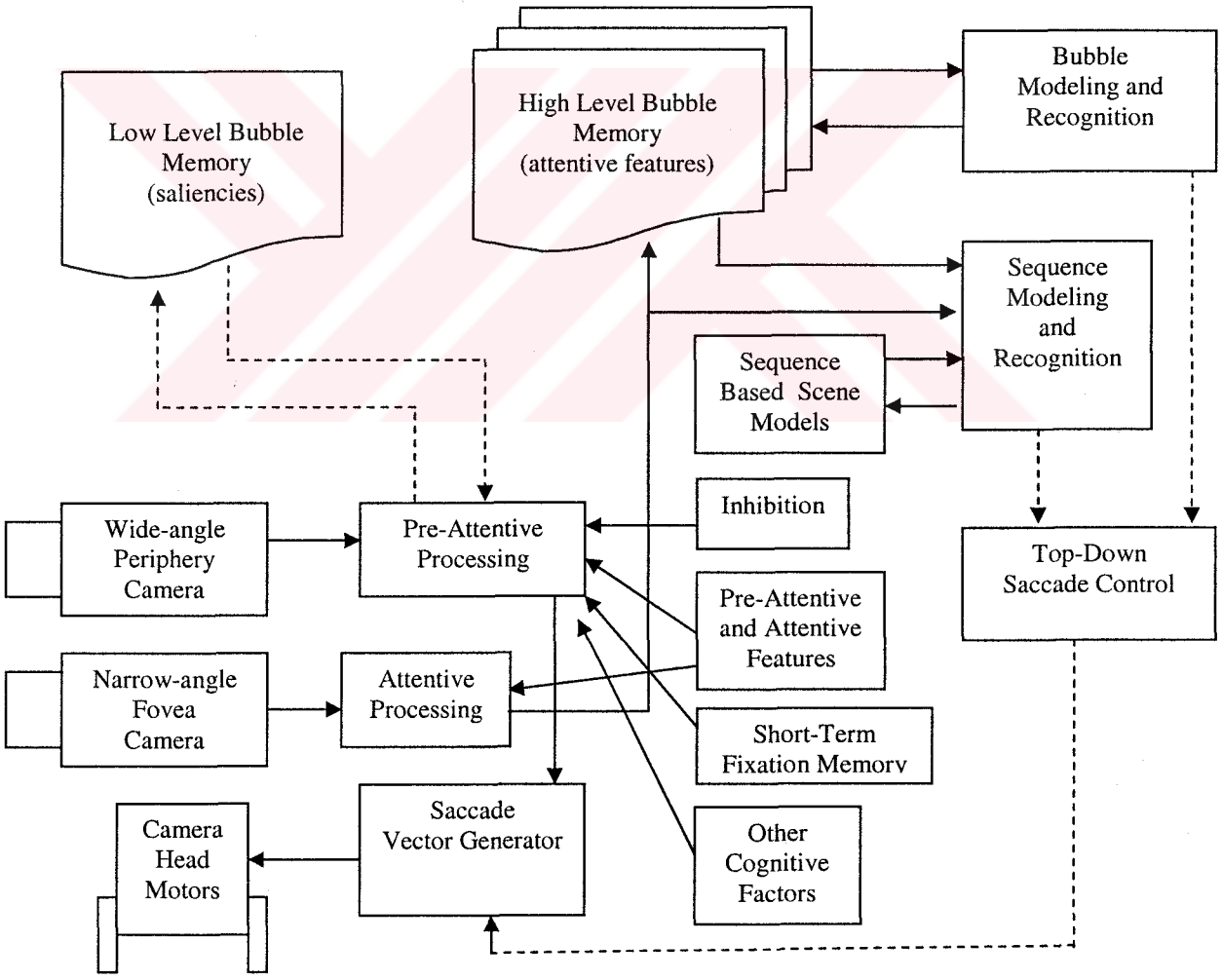

Figure 6.1 An integrated model of attentive vision.

In Figure 6.1 the periphery and fovea images are obtained by the two-camera sensor system simulating the human retina. Then the periphery image is fed into the pre-attention system, which also receives pre-attentive interest criteria, inhibition settings, fixation memory contents and any other higher level cognitive effects like directional selectivity. Results of pre-attentive processing are saved in a bubble memory, so that they can be recalled when part of the same peripheral field needs to be processed at a later time, or when visual information is not available. The new fixation point is selected and sent to the saccade controller which generates the saccade vector and commands motor units. In humans saccades are also known to be controlled in a predictive manner, based on expectations about a scene. This top-down saccade control mechanism, which is not tested on APES, has inputs from sequence processing and bubble memory units as shown by dashed lines.

At each fixation the fixation fovea is processed to extract attentive features. Similar to pre-attentive interest criteria, the system also enables any feature that can be computed on the fovea image to be used as an attentive feature. These features are sequentially processed by attentional sequence modeling and recognition algorithms, which enable temporal recognition. Attentive features are also stored in a bubble memory to form an higher-level, viewpoint dependent, visual model of the environment.

7. CONCLUSION

This thesis presents a model of attentive vision which is then implemented on a mobile robot – APES: The model is based on the key properties of biological vision - which has been suggested by neurophysiological and psychological studies on biological systems. In particular, the following integral characteristics are taken into consideration: Fovea - periphery distinction in the retina, selective attention, inhibition of return and dead zone during attention, temporal recognition, and visual integration over saccades.

The two-camera sensor configuration provides a realistic retina model by changing photoreceptor densities in foveal and peripheral regions. Similarly the fixation memory and inhibition features embedded in the attention system simulate the inhibition of return and dead zone mechanisms respectively, resulting in an intelligent looking behavior. The behavior of the attention system can also be controlled by changing its attention criteria based on the given task. Such behavior is also clearly demonstrated by humans, although the underlying mechanisms are unknown. Temporal recognition is achieved by probabilistic and evidence based algorithms, and the bubble memory is a model for visual integration.

This integrated model of active and attentive vision has enabled the following contributions:

1- The use of attentional sequences in scene modelling and recognition: The attentional sequence is the essential output of any attentive vision system, regardless of its attention mechanism or other internal properties. Our methods based on Markov models and Dempster-Shafer theory of evidence provide two such basic tools for modeling and classification of spatio-temporal attentional sequences. Our experiments show that both sequence classification methods work well, with evidential reasoning having practical advantages over Markov models. The theory of evidence is also more compatible with the general idea of collecting information in time to reach a decision, which is the basic assumption of temporal vision. We have also shown that scene recognition using

attentional sequences is a different problem. The ability to model and recognize attentional sequences successfully does not necessarily guarantee good scene or object classification performance. Besides sequence recognition, scene classification performance is largely dependent on the feature space and the quality of learning sessions. However, we have shown that even with a simple edge based feature space, simple shapes were correctly classified and satisfactory results were obtained on complex real world scenes depending on learning performance.

2. A model for visual integration over saccades: As a model for vision based environment representation we propose the bubble memory, inspired by the theory of visual integrative buffer in cognitive psychology. The basic assumption of the bubble approach is the use of visual information only, to create a practically useful environment model. The results of visual computations are mapped on to a spherical surface in a subject centered, viewpoint dependent spherical coordinate system. This is thought to be the most natural representation of visual information collected by a moving active and attentive vision system. In our experiments we demonstrated that the bubble representation is useful for both environment modeling and scene recognition.

3. An integrated model of visual attention: Our attention model simulates some of the key properties of biological vision described above and enables an intelligent looking behavior

4. Implementation by a robotic system: Our mobile robot APES, which was originally developed for active vision research, has gone through considerable modifications in order to be able to embed the proposed model as the basis of its visual system. Its sensor system has been extended to a two-camera configuration, its motion control hardware and stability is improved to achieve faster and more accurate saccades, and finally its software is completely changed to implement the developed algorithms.

Our work on attentional sequences and the bubble model has also raised some interesting questions about scene perception, which require further investigation. For example, metrics for defining actual partitions in the sequence space or the representation capability of a given sequence for a given scene are two interesting topics, shortly mentioned in this thesis. The use of bubbles as a scene recognition tool is already

demonstrated. However the viewpoint dependent integrative nature of bubbles also make them an interesting long term memory mechanism, especially for moving vision systems. Humans' ability to perform tasks based on previously obtained visual information, like moving in a well known dark room, can be demonstrated by such memories. Also, the computation of a true 3D environment model using bubbles generated at different viewpoints may be possible. The performance of these algorithms on robotic systems can be compared to actual human performance. We believe that joint modeling and implementation efforts by biological and computational vision communities can potentially generate new questions and answers in vision, and lead to success in both directions. We consider our current effort as a contribution in this respect.

REFERENCES

1. Goodale, M.A.. "The Cortical Organization of Visual Perception and Visiomotor Control", in S.M.Kosslyn, D.N. Osherson, editors, *Visual Cognition*, pp. 167-214. MIT Press, 1995.
2. Gallant, J.L., D.C.Van Essen and H.C. Nothdurft, "Two-Dimensional and Three Dimensional Texture Processing in Visual Cortex of the Macaque Monkey", in T.V. Papathomas, C. Chubb, A. Gorea and E. Kowler, editors, *Early Vision and Beyond*, pp. 89-98, MIT Press, 1995.
3. Gallant, J. L., C.E. Connor, S. Rakshit, J.W. Lewis and D.C. Van Essen, "Neural Responses to Polar, Hyperbolic and Cartesian Gratings in Area V4 of the Macaque Monkey", *Journal of Neurophysiology*, Vol.76, No. 4, pp. 2718-2739, 1996.
4. Hubel, D.H., "Eye, brain and vision", *Scientific American Lib.*, 1988.
5. Marr, D. and E. Hildreth, "Theory of Edge Detection", *Proceedings of Royal Society of London*, B207, 187-217, 1980.
6. Julesz, B., *Dialogues on Perception*, MIT Book Press, Cambridge, MA 1995.
7. Kandel, E.R. and J.H.Schwartz, editors, *Principles of Neural Science*, Elsevier, 1986.
8. Malik, J. and P. Perona, "Preattentive Texture Discrimination With Early Vision Mechanisms", *Journal of Opt. Soc. America*, A7, pp. 923-932, 1990.
9. Sagi, D., "The Psychophysics of Texture Segmentation", in T.V. Papathomas, C. Chubb, A. Gorea and E. Kowler, editors, *Early Vision and Beyond*, pp. 69-77, MIT Press, 1995.
10. Barrow, H. and J.M. Tenenbaum, "Computational Vision", *Proceedings of IEEE*, 69:572-575, 1981.
11. Caelli, T., "A Brief Overview of Texture Processing in Machine Vision", in T.V. Papathomas, C. Chubb, A. Gorea and E. Kowler, editors, *Early Vision and Beyond*, pp. 89-98, MIT Press, 1995.
12. Marr, D., *Vision*, W.H.Freeman, 1982.
13. Ballard, D.H. and Brown, C.M. *Computer Vision*. Prentice-Hall, 1982.
14. Zeki, S., "The Visual Image in Mind and Brain", *Scientific American*, Vol.267, No.3, September 1992.

15. Brooks, R.A., A Robust Layered Control System for a Mobile Robot, *IEEE Journal of Robotics and Automation*, 2: 109-126, 1986.
16. Haralick, R. and L. Shapiro, *Computer Vision*, Prentice-Hall, 1985.
17. Widrow, B. and M. Lehr, "30 Years of adaptive Neural Networks: Perceptron, Madaline, and Backpropagation", *Proceedings of IEEE*, Vol.78, No.9, September 1990.
18. Kowler, E., editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, Elsevier, 1990.
19. Kowler, E, "Eye Movements", in S.M.Kosslyn, D.N. Osherson, editors, *Visual Cognition*, pp. 215-266, MIT Press, 1995.
20. Noton, D. and L. Stark, "Scanpaths in Eye Movements During Pattern Recognition", *Science*, Vol.171, pp. 308-311, January 1971.
21. Stark, L. and S.R. Ellis, "Scanpaths Revisited: Cognitive Models Direct Active Looking", in Fisher, Monty and Senders, editors, *Eye Movements: Cognition and Visual Perception*, pp. 193-226, Erlbaum, NJ, 1981.
22. Gouras, P. and C.H.Bailey, "The Retina and Phototransduction", in J.H.Schwartz and E.R.Kandel, editors, *Principles of Neural Science*, Elsevier, 1986.
23. Clark, J., "Spatial attention and Latencies in Saccadic Eye Movements", *Vision Research* Vol. 39 pp. 585-602, 1999.
24. Dosher,B.A. and Z. Lu., "Mechanisms of Perceptual Attention in Precuing of Location", *Vision Research*, 40:1269-1292, 2000.
25. Gouras, P., "Oculomotor System", in J.H.Schwartz and E.R.Kandel, editors, *Principles of Neural Science*, Elsevier, 1986.
26. Greene,H.H., "Temporal Relationships Between Eye Fixations and Manual Reactions in Visual Search", *Acta Psychologica*, 101:105-123, 1999.
27. Henderson, J.M., "Visual attention and the attention-action interface", in K.Akins, editor, *Perception*, pp. 290-316, Oxford University Press, 1996.
28. Kowler, E, "The Role of Visual and Cognitive Processes in the Control of Eye Movements", in E.Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 1-63, Elsevier, 1990.
29. Liversedge, S.P. and J.M. Findlay, "Saccadic Eye Movements and Cognition",
30. Pavel, M., "Predictive Control of Eye Movement", in E.Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 71-112, Elsevier, 1990.

31. Viviani, P., "Eye Movements in Visual Search: Cognitive, Perceptual and Motor Control Aspects", in E.Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 71-112, Elsevier, 1990.
32. Wolfe, J.M., "What Can 1,000,000 Trials Tell Us About Visual Search", *Psychological Science*, Vol.9, No.1, January 1998.
33. Zingale, C.M. and E. Kowler, "Planning Sequences of Saccades", *Vision Research*, Vol.27, No.8, pp. 1327-1341, 1987.
34. Parkin, A.J., *Essential Cognitive Psychology*, Psychology Press, 2000.
35. Abbott, A.L et al, "Promising Directions in Active Vision", *International Journal of Computer Vision*, 11:2, 109-126, 1993.
36. Aloimonos, J., "Purposive and Qualitative Active Vision", *Proceedings of Image Understanding Workshop*, September 1990.
37. Ballard, D.H., "Animate Vision", *Artificial Intelligence*, 48: 57-86, 1991.
38. Ballard, D.H. and C.M.Brown, "Principles of Animate Vision", *CVIP: Image Understanding*, 56(1), July 1992.
39. Itti, L. and C. Koch, "Coputational Modelling of Visual Attention", *Nature Reviews*, Vol.2, February 2001.
40. Koch, C. and S. Ullman, "Selecting One Among the Many: A Simple Network Implementing Shifts in Selective Visual Attention", *A.I. Memo* 770, C.B.I.P Paper 003, 1994.
41. Li,G. and B. Svensson, "Navigating With a Focus-Directed Mapping Network", *Autonomous Robots*, 7:9-30, 1999.
42. Maris, M, "Attention-Based Navigation in Mobile Robots Using a Reconfigurable Sensor", *Robotics and Autonomous Systems*, 34:53-63, 2001.
43. Soyer, Ç. H.I.Bozma, and Y. Istefanopulos, "A Mobile Robot with a Biologically Motivated Active Vision System", *Proceedings of IEEE RSC International Conference on Intelligent Robots and Systems*, IROS, 1996.
44. Westin, C. et al, "Attention Control for Robot Vision", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 726-1996.
45. Kandel, E.R., J.H. Schwartz, and T.M. Jessell, editors, *Essentials of Neural Science and Behavior*, Appleton & Lange, 1995.

46. Sagi, D., "The Psychophysics of Texture Segmentation", in T.V. Papathomas, C. Chubb, A. Gorea and E. Kowler, editors, *Early Vision and Beyond*, pp. 69-77, MIT Press, 1995.
47. Kelly, J.P., "Anatomy of central visual pathways", in J.H.Schwartz, J.H. and E.R.Kandel, editors, *Principles of Neural Science*, Elsevier, 1986.
48. Viviani, P., "Eye Movements in Visual Search: Cognitive, Perceptual and Motor Control Aspects", in E.Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 71-112, Elsevier, 1990.
49. Connor, C.E., D.C. Preddie, J.L. Gallant and D.C. Van Essen, "Spatial Attention Effects in Macaque Area V4", *The Journal of Neuroscience*, 17(9) pp. 3201-3214, 1997.
50. Horowitz,T.S. and J.M. Wolfe, "Visual search has no memory", *Nature*, Vol.357, pp. 575-577, August 6, 1998.
51. Itti,L. and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention", *Vision Research*, 40:1489-1506, 2000.
52. Kapoula,Z. and D.A. Robinson, "Saccadic Undershoot is not Inevitable: Saccades Can Be Accurate", *Vision Research*, Vol.26, No.5, pp. 735-743, 1986.
53. Tagare, H. And K. Toyama and J. G. Wang, "A Maximum Likelihood Strategy for Directing Attention During Visual Search", *IEEE Transactions on PAMI*, Vol.23, No 5, pp. 491-500, May 2001.
54. McGaugh, J.L., N.M.Weinberger and G.Lynch editors. *Brain and Memory*, Oxford University Press, 1995.
55. Pavel, M., "Predictive Control of Eye Movement", in E.Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 71-112, Elsevier, 1990.
56. Rimey, R.D. and C.M.Brown, "Selective Attention as Sequential Behaviour: Modelling Eye Movements with an Augmented Hidden Markov Model", Technical Report, The University of Rochester, Computer Science Department, February 1990.
57. Rimey, R.D. and C. Brown, "Control of Selective Perception Using Bayes Nets and Decision Theory", *International Journal of Computer Vision*, 12:2/3, 173-207, 1994.
58. Rensink,R.A., "Seeing, Sensing, and Scrutinizing", *Vision Research*, 40:1469-1487, 2000.
59. Sandini, G., F.Gandolfo, E.Grosso, and M.Tistarelli, "Vision During Action", in Y.Aloimonos, editor, *Active Perception*, Lawrence Erlbaum Associates, 1993.

60. Schlingensiepen, K.H., et al, "The Importance of Eye Movements in the Analysis of Simple Patterns", *Vision Research*, Vol.26, No.7, pp. 1111-1117, 1986.
61. Horowitz,T.S. and J.M. Wolfe, "Visual search has no memory", *Nature*, Vol.357, pp. 575-577, August 6, 1998.
62. Malinov, I.V. and J. Epelboim, A.N. Herst, R.M. Steinman, "Characteristics of Saccades and Vergence in Two Kinds of Sequential Looking Tasks", *Vision Research* V.40, pp. 2083-2090, 2000.
63. Julesz, B., *Dialogues on Perception*, MIT Book Press, Cambridge, MA 1995.
64. McGaugh, J.L., N.M. Weinberger and G. Lynch, editors, *Brain and Memory*, Oxford University Press, 1995.
65. Rybak, I. A., V. I. Gusakova, A. V. Golovan, L.N. Podladchikova and N. A. Shevtsova, "A Model of Attention-Guided Visual Perception and Recognition", *Vision Research*, Special Issue: Models of Recognition, 1998.
66. Fiala,J.C. et al, "TRICLOPS: A Tool for Studying Active Vision", *International Journal of Computer Vision*, 12:2/3, 231-250, 1994.
67. Papanikolopoulos, N.P., "Adaptive Control, Visual Servoing, and Controlled Active Vision", *Proceedings of IEEE International Conference on Robotics and Automation*, 1994.
68. Rao, R.P.N. et al, "Modeling Saccadic Tergeting in Visual Search", in Touretzky, D., Mozer, M. and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8 (NIPS*95)*, MIT Press, 1996.
69. Treisman, A., and G. Gelade, "A Feature Integration Theory of Attention", *Cognitive Psychology*, 12, pp. 97-136, 1980.
70. Tsotsos, J.K. et al, "Modeling Visual Attention via Selective Tuning", *Artificial Intelligence*, 78:507-545, 1995.
71. Ballard, D.H., "On the Function of Visual Representations", in K.Akins, editor, *Perception*, pp. 111-131. Oxford University Press, 1996.
72. Lago-Fernandez, L.F., M.A. Sanchez-Montanes, and F. Cobacho, "A Biologically Inspired Visual System for an Autonomous Robot", *Neurocomputing*, 38-40:1385-1391, 2001.
73. Bozma, H.I. and Ç.Soyer, "Shape Identification Using Probabilistic Models of Attentional Sequences", *Proceedings of Workshop on Machine Vision Applications*, IAPR, 1994.

74. Wasson,G., D. Kortenkamp and E. Huber, "Integrating Active Perception with an Autonomous Robot Architecture", *Robotics and Autonomous Systems*, 29:175-186, 1999.

75. Soyer, Ç. H.I.Bozma, and Y. Istefanopulos, "A Mobile Robot with a Biologically Motivated Active Vision System", *Proceedings of IEEE RSC International Conference on Intelligent Robots and Systems*, IROS, 1996.

76. Soyer, Ç. and H.I.Bozma, "Further Experiments in Classification of Attentional Sequences: Combining Instantaneous and Temporal Evidence", *Proceedings of IEEE 8th International Conference on Advanced Robotics*, ICAR, 1997.

77. Soyer, Ç., "A Mobile Robot with a Biologically Motivated Vision System", MS Thesis, Boğaziçi University, 1995.

78. Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, 77(2), February 1989.

79. Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

80. Breitmeyer, B.G., *Visual Masking: An Integrative Approach*, Oxford University Press, 1984.

81. Bridgeman, B., A. van de Heijden, and B. Velichkovsky, "A Theory of Visual Stability Across Saccadic Eye Movements", *Behavioral and Brain Sciences, 17*, 247-292, 1994.

82. McConkie, G.W. and C. Currie, "Visual Stability Across Saccades While Viewing Complex Pictures", *Journal of Experimental Psychology: Human Perception and Performance, 22(3)*, 563-581, 1996.

VITA

Çağatay Soyer was born in İzmit, Turkey, on May 13th, 1970. After graduating from Kadıköy Anadolu High School in 1988 he studied Mechanical Engineering at Istanbul Technical University. In 1990 he left this university and started studying Electrical and Electronic Engineering at Boğaziçi University, Istanbul. He received his BS degree in Electrical and Electronic Engineering in 1993 and MS degree in Biomedical Engineering in 1995 both from Boğaziçi University.

The PhD study presented in this thesis started in October 1995 and was completed in January 2002 at the Institute of Biomedical Engineering, Boğaziçi University. It was supported by TUBITAK, Boğaziçi University Research Fund and Boğaziçi University Foundation. Four papers out of this work were presented at international conferences and one paper was accepted for publication in *IEEE Transactions on Systems, Man, and Cybernetics*.