Attentional Sequence-Based Recognition: Markovian and Evidential Reasoning

Çağatay Soyer, H. Işıl Bozma, and Yorgo İstefanopulos, *Member, IEEE*

***Abstract*—Biological vision systems explore their environment via allocating their visual resources to only the interesting parts of a scene. This is achieved by a selective attention mechanism that controls eye movements. The data thus generated is a sequence of subimages of different locations and thus a sequence of features extracted from those images—referred to as attentional sequence. In higher level visual processing leading to scene cognition, it is hypothesized that the information contained in attentional sequences are combined and utilized by special mechanisms—although still poorly understood. However, developing models of such mechanisms prove out to be crucial—if we are to understand and mimic this behavior in robotic systems. In this paper, we consider the recognition problem and present two approaches to using attentional sequences for recognition: Markovian and evidential reasoning. Experimental results with our mobile robot APES reveal that simple shapes can be modeled and recognized by these methods—using as few as ten fixations and very simple features. For more complex scenes, longer attentional sequences or more sophisticated features may be required for cognition.**

***Index Terms*—Active vision, attentional sequence classification, Dempster–Shafer theory, Markov models, selective attention.**

I. Introduction

BIOLOGICAL vision systems have the capability of allocating their visual resources to different parts of a scene in time by shifting their attention [29], [32], [33], [36], [46]. 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 [10], [11], [30], [31], [33], [36], [47]. The incoming stream of subimages is then utilized to generate a related sequence of features extracted from different spatial locations at different times, referred to as the 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 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. There has been a lot of work in vision science and active vision on the generation of attentional sequences. However, the next step of linking the attentional sequences to visual tasks and the responsible higher level mechanisms are poorly understood. Yet, developing models of such mechanisms prove out to be crucial—if robotic vision systems are to make use of attentional sequences—as they need to do in many applications [1], [2], [4], [5], [39], [42]. In this paper, we consider two-dimensional (2-D) recognition tasks and propose two approaches regarding how to use attentional sequences for 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. Second, the approaches are capable of handling variations with regards to scanpaths taken. To understand the implication of this, consider the possibility of looking at a rectangular blob. Then, we might look at each edge consecutively or we may first look at the top and bottom edges and then the left and right edges. Thus, even with a single scene, the sequence of observed features may change depending on the scanpath taken. Finally, these approaches have mechanisms for learning under external supervision. These approaches have been implemented in our mobile robot APES, which has both physical and mental attention capabilities.

In the remainder of this section, we present a very brief overview of the physiology of human vision in order to point out the key features integral to biological vision and discuss how work on active vision has made use of it. In the next section, two approaches to using attentional sequences, Markov and evidential models, are presented. We then describe APES, its vision system, and selective attention mechanism. Comparative experimental results performed on simple and complex scenes demonstrate the efficacy of these approaches.

A. Biological Visual Attention

Physiological and psychological studies suggest four key properties of biological vision: Fovea-periphery distinction on the retina, occulomotion, image representation, and serial processing. First, biological vision systems process only a small part of their visual field in detail. Unlike traditional cameras, the distribution of receptor cells on the retina is like a Gaussian with a small variance, resulting in a loss of resolution as we move away from the optical axis of the eye [11], [31]. The small region of highest acuity around the optical axis is called the fovea, and the rest of the retina is called periphery. Second, as a consequence of this fovea-periphery distinction,

Manuscript received September 4, 2001; revised February 9, 2002. This work was supported by Bogazici University under Research Fund Grant 01A202D. This paper was recommended by Associate Editor D. Goldgof.

Ç. Soyer and Y. İstefanopulos are with the Institute of Biomedical Engineering, Boğaziçi University, Bebek, İstanbul 81030, Turkey (e-mail: soyer@boun.edu.tr; istef@boun.edu.tr).

H. I. Bozma is with the Department of Electrical and Electronic Engineering, Boğaziçi University, Bebek, İstanbul 81030, Turkey (e-mail: bozma@boun.edu.tr).

Digital Object Identifier 10.1109/TSMCB.2003.810904

saccades—very rapid jumps of optical axis—are used to bring images of chosen objects to fovea where resolution of fine visual detail is at its best [10], [33]. Saccadic eye movements are one of the capabilities of the oculomotor system which is involved in controlling the eye movements. Saccades take place at very high speeds up to 600°–700°/s and there is a delay of about 0.2 s between detecting a target and making a saccade. Saccadic eye movements 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 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 [9], [23], [24], [28], [30], [31], [47]. This finding implies that image representations in higher visual centers are related to levels of abstraction. 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. In other cortical regions, collectively called the visual association cortex or prestriate cortex, cells respond to color, motion, and orientation. Finally, 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 [12], [13], [15], [29]. 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. Thus, approaches to using the attentional sequence thus generated must be developed.

B. Relation to Previous Work

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 [1], [2]. This is done by first implementing a simple 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 to attentive processing. Active vision research has mostly concentrated on generating fixations and controlling camera movements [1], [4], [41], [42]. 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 [12], [13]. An alternative approach based on simplified visual search mechanisms such as using attractive forces has been presented in [48] and [49]. 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 [39]. A generalization of these ideas to Bayes networks and decision theory is presented in [40]. A maximum-likelihood strategy for directing attention has been applied in recognition tasks [50]. All of these approaches have provided different mechanisms of fixation generation by specifying a set of points on the image and work well in many applications. The use of resulting attentional sequences for making decisions about the current task remained relatively unexplored [7], [45].

Our aim is quite different from these studies. We do not want to implement a specific attentional sequence generating mechanism. Rather, we ask how to use the data thus collected for recognition purposes. In particular, we are interested in reasoning that is independent of the pre-attentive and attentive cues used.

C. Mathematical Formulation: Pre-Attention and Attention

We assume that the visual processing is composed of a loop of pre-attentive and attentive stages which generate an attentional sequence.

In the pre-attention stage, 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^f 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 $I_f^c \in (I_v^t)$ 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

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

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 Tth fixation.

In the attentive stage, each fovea T_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 mth 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 $\Omega \triangleq \Omega_1 \times \cdots \times \Omega_M$. Note that

$$|\Omega| = \prod_{m=1}^{M} N_m. \tag{2}$$

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

$$o^t = \left[f_1 \left[I_f^t \right], \ldots, f_M \left[I_f^t \right] \right]. \tag{3}$$

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 set of values of vi-

sual primitives observed at different locations at different times, 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.

D. Problem Statement

Suppose that the visual task is defined as follows. The vision system is looking at a scene in an attentive manner and thus generating an attentional sequence O^T. Furthermore, the system knows about L different scenes to which the scene currently being looked could or could not belong. Then find $1^* \in L$ that best explains the observed attentional sequence O^T.

II. Using Attentional Sequences

A. Markov Models and Reasoning

In this approach, the attentional sequence $O^T = (o^l, \ldots, o^T)$ is considered as a discrete Markov process [38] with an alphabet Ω. This process is associated with the transition probability matrix A of dimension $|\Omega| \times |\Omega|$

$$A = \{P(o^{t+1} = v^j \mid o^t = v^i)\} = \{a_{ij}\}$$

where

$$v^i, v^j \in \Omega$$

and

$$\sum_{j \in \Omega} a_{ij} = 1, \quad \forall i \in \Omega. \tag{4}$$

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 objects or scenes, 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 III-C.

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

$$P(O^T \mid l) = P(o^1) \cdot \prod_{t=1}^{T-1} P(o^{t+1} \mid o^t, l), \quad \text{where } P(o^1) = \frac{1}{|\Omega|}. \tag{5}$$

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

$$l^* = \arg\max_{l \in L} P(O^T \mid l). \tag{6}$$

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 [38].

B. Evidential Models and Reasoning

In this approach, the attentional sequence $O^T = (o^l, \ldots, o^T)$ 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 [43], [45]. 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 1-1 correspondance with the subsets 2^L of L. Thus, we use $A \in 2^{\mathcal{L}}$ to denote a proposition. In classification, we are in particular interested in propositions of the form

$$A_l = \{l\}, \quad l = 1, \ldots, L \qquad \text{where } L = |\mathcal{L}|. \tag{7}$$

Now suppose for each proposition A_l, we have a transition frequency matrix $T_l : \Omega \times \Omega \to [0, \infty]$. 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 $\omega : 2^L \times \Omega \to [0, \infty]$ represent the weight of evidence function. Then

$$\omega(A_l, o^t) = T_l(o^{t-1}, o^t). \tag{8}$$

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 $s_l : 2^L \times \Omega \to [0, 1]$ define a simple support function focused on A_l. Then s_l can be defined as

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

where

$$s_l(A_l, o^t) = 1 - e^{-c\omega(A_l, o^t)}. \tag{9}$$

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

However, each evidence points to a set of propositions $A_l, l = 1, \ldots, L$ with different degrees of support $s_l(A_l, o^t)$. Since $A_l \cap A_k = \emptyset$, 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_l focused on A_l are given with basic probability numbers (10)–(12), shown at the bottom of the page.

The effect of s_l^i is to provide instantaneous support for each proposition A_l. 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 $s_l^t : 2^L \times \Omega^t \rightarrow [0,1]$ 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 A_l with $s_l^i(A_l)$ and s_1^t focused on A_l with $s_l^t(A_l)$. The cumulative support $s_l^{t+1} : 2^L \times \Omega^{t+1} \rightarrow [0,1]$ is defined recursively as the orthogonal sum $s_l^{t+1} = s_l^t \oplus s_l^i$ (see (13), shown at the bottom of the page).

Then, the result of classification is given by

$$l* = \arg\max_{l \in L} s_l^{t+1}(A_l, O^{t+1}). \tag{14}$$

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.

C. 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. We assume that the vital processing is composed of a loop of pre-attentive and attentive stages which generated an attentional sequence, as shown in Fig. 1. 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 one. 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 nonexisting 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.

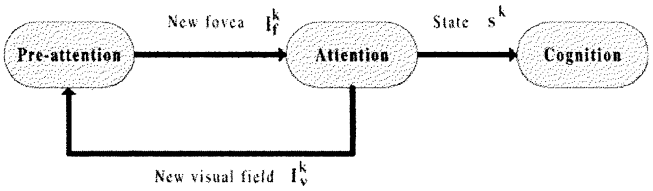


Fig. 1. General flow of processing.

III. APES—Active Perception System

APES is a simple mobile robot with an active vision system [44], as shown in Fig. 2. Its body is a mobile vehicle with two driven conventional wheels, and one freely rotating support wheel. 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

$$m(A_l, o^t) = \frac{s_l(A_l, o^t) \prod_{\substack{j=1 \\ j \neq l}}^{L} (1 - s_j(A_j, o^t))}{1 - \prod_{j=1}^{L} s_j(A_j, o^t)} \tag{10}$$

$$m(L, o^t) = \frac{\prod_{j=1}^{L} (1 - s_j(A_j, o^t))}{1 - \prod_{j=1}^{L} s_j(A_j, o^t)} \tag{11}$$

$$s_l^i(C, o^t) = \begin{cases} 0, & \text{if } C \text{ contains none of } A_l, l = 1, \ldots, L \\ \dfrac{s_l(A_l, o^t) \prod_{\substack{j=1 \\ j \neq l}}^{L} (1 - s_j(A_j, o^t))}{1 - \prod_{j=1}^{L} s_j(A_j, o^t)}, & \text{if } C \text{ contains } A_l \text{ but does not contain } A_j, \\ & j = 1, \ldots, L, j \neq l \\ \dfrac{\sum_{C \cap L} s_l(A_l, o^t) \prod_{\substack{j=1 \\ j \neq k}}^{L} (1 - s_j(A_j, o^t))}{1 - \prod_{j=1}^{L} s_j(A_j, o^t)}, & \text{if } C \text{ contains some of } A_l, C \neq L \\ 1, & \text{if } C = L. \end{cases} \tag{12}$$

$$s_l^{t+1}(C, O^{t+1}) = \begin{cases} 0, & \text{if } C \text{ does not contain } A_l \\ 1 - \left(1 - s_l^i(A_l, o^{t+1})\right)\left(1 - s_l^t(A_l, O^t)\right), & \text{if } C \text{ contains } A_l \\ 1, & \text{if } C = L. \end{cases} \tag{13}$$

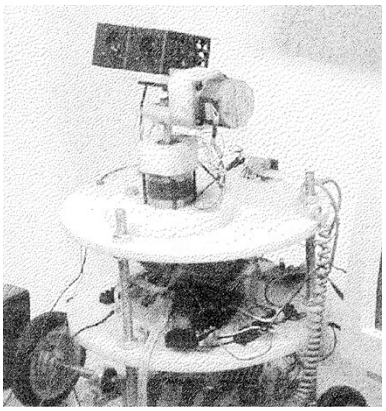


Fig. 2. APES robot and its 2-DOF camera base.

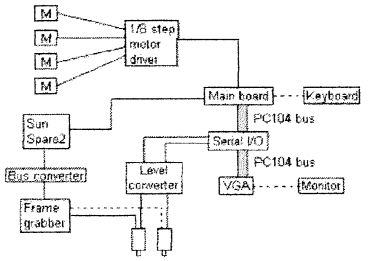


Fig. 3. Schematic of APES.

TABLE I
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

designed to be cocentered, 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 a research platform where different selective attention algorithms are implemented and used.

Fig. 3 and Table 1 show the hardware configuration, and Fig. 4 shows the simplified selective attention mechanism and basic vision process, respectively. 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. During operation new fixation point in the periphery is determined by visual processing and 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.

The vision system of APES is inspired by the key properties of biological vision as follows.

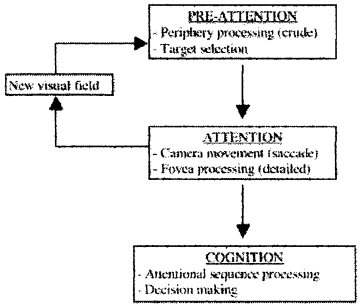


Fig. 4. Simplified selective attention mechanism of APES.

1) *Fovea–Periphery Distinction:* APES can simulate the nonuniform photoreceptor density of the human retina in three ways:
   a) by varying complexity of features extracted from foveal and peripheral regions;
   b) by processing foveal and peripheral images at different resolutions;
   c) by using a two-camera retina model, which uses separate camera angles and pixel densities for foveal and peripheral images.

   In this paper, the first retina model is used, as explained in Sections II-A and II-B.
2) *Oculomotion:* The two degrees of freedom (DOF) step motor-based head assembly and the motion of the camera cannot be compared to the highly developed and poorly understood occulomotor system of mammals. However, APES can effectively control the optical axis of the camera and the fixation point with an accuracy of 1.8° 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.
3) *Levels of Representation:* The attention criteria (also called salient features) guide peripheral processing and is used to determine the next fixation point. APES can use either edge content (computed by the gradient) or brightness. Currently, we are working on enrichening this set to include Cartesian and non-Cartesian filters.
4) *Serial Processing:* Finally, the selective attention mechanism employed by APES guarantees that only the most important parts of a scene are fixated and processed in detail, and relevant information is collected and integrated in time to solve the given task.

A. APES's Active Vision

Within this framework, visual processing consists of three basic stages of operation: pre-attention, attention, and cognition, as shown in Fig. 4. The visual field components are shown in Fig. 5. APES finds a new fovea by considering overlapping candidate foveas within its 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, APES has two mechanisms—inhibition and memory—that get activated before a saccade is made in order to avoid processing the same areas twice or going into infinite fixation loops. These are motivated

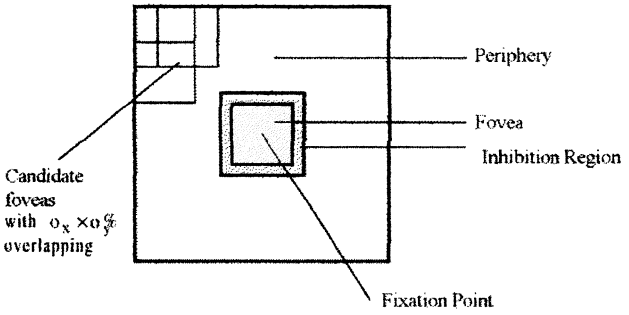


Fig. 5. APES's visual field and its components.

by vision science findings that indicate the presence of spatial inhibition mechanisms that delay fixations on an area that has just been fixated.

The inhibition mechanism works as follows: An $H \times H$ pixel region I_h^t around the currently fixated fovea I_f^t, at the center of the current visual field I_v^t, is defined as the inhibition region. During pre-attentive processing, all candidate foveas $I_f^c \in C(I_h^t)$ falling within the inhibition region are inhibited and cannot be chosen as the next fixation fovea. In this manner, the inhibition mechanism also enables the control of saccade magnitudes.

The memory mechanism works via keeping track of previously fixated foveas in terms of camera joint angles and preventing fixations on these targets even if they are not within the current inhibition region. Since APES has finite precision in pan and tilt directions, $\Delta\theta = \Delta\varphi = 1.8°$, there is a finite set of fixation points in camera coordinates, given by $\beta = \{(\theta, \varphi) \in \Re^2 \,|\, \theta \equiv i \cdot \Delta\theta, \varphi \equiv j \cdot \Delta\varphi\}$, where i and j are integers.

In order to keep track of previously fixated coordinates (θ, φ), we use a first-in-first-out memory of size D. Then, the corresponding set of foveas is $C_d = \{I_f^t, I_f^{t-1}, \ldots, I_f^{t-D}\}$. 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

$$\tilde{a}\left(I_f^c\right) = \begin{cases} 0, & \text{if } I_f^c \in C\left(I_h^t\right) \\ 0, & \text{if } I_f^c \in C_d \\ a\left(I_f^c\right), & \text{if } I_f^c \in C\left(I_v^t\right), I_f^c \notin C\left(I_h^t\right), C_d. \end{cases} \tag{15}$$

Note that APES can use any simple image feature as low-level attention criteria 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 this paper, $a(I_f^c) \triangleq \sum_{p \in I_f^c} |\nabla I(p)|$.

In the attentive stage the fixation fovea is subjected to more detailed processing. APES can extract various complex features 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. In the experiments reported in this paper, a very simple feature set $\Omega = \Omega_1$ is considered. The set Ω_1 is defined as $\Omega_1 \triangleq \{i \,|\, i = 0, \ldots, 7\}$ where each value $i = 0, 1, 2, 3$ indicates an edge oriented $i \times 90°$ and each value $i = 4, 5, 6, 7$ indicates an edge oriented $i \times 90° + 45°$.

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° rotated versions of the same edge, therefore rotation invariance can only be expected up to $\pm$ 22.5° even if edge detection is noiseless.

Each observation o_t is added to the attentional sequence $O^T = (o^1, \ldots, o^T)$, thus generating a data sequence. The cognitive stage works with the incoming 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. In the next section we introduce mathematical methods developed for modeling and recognition of attentional sequences.

IV. 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 modeling on classification performance.

In these experiments, APES used a 200×200 pixel visual field and a 40×40 pixel fovea. The overlap between candidate foveas was 50% and a fixation memory depth $D = 10$ is used to inhibit the last ten 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 II-B. In the attentive stage, the feature space consists of $\Omega = \Omega_1$ corresponding to eight 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 3×3 operator for detecting edges with an orientation of $i°$. 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 the following:

1) slight variations in the first fixation point;
2) small positioning errors in the camera head assembly;

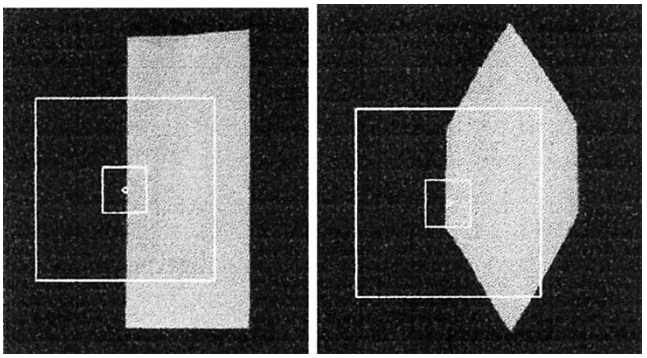


Fig. 6. Simple scenes containing "rectangle" and "polygon."

	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

Fig. 7. Scene 1: Learning using attentional 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

Fig. 8. Scene 2: Learning using attentional sequences of length 10.

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.

A. Simple Scenes

The first set of experiments was performed on simple 2-D shapes hanging on a black background, as shown in Fig. 6. 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.

In the first set of experiments, sequences are of length 10. The observed feature transition frequencies are shown in Figs. 7 and 8. Even with attentional sequences of length 10, these matrices start to become differentiable. The matrix for Scene 1 favors no transitions between diagonal features 4–6, and 7, as compared

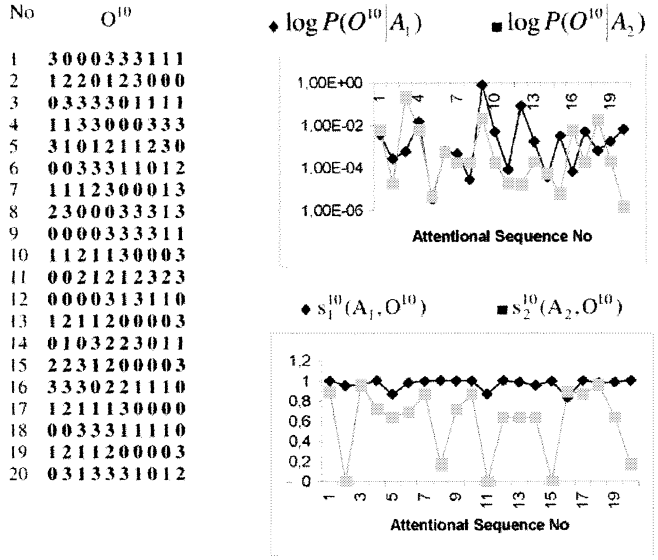


Fig. 9. Results after ten fixations on Scene 1 with ten fixation learning on Scene 1 and Scene 2. Recognition rate is 65% with Markov models and 90% with evidential reasoning.

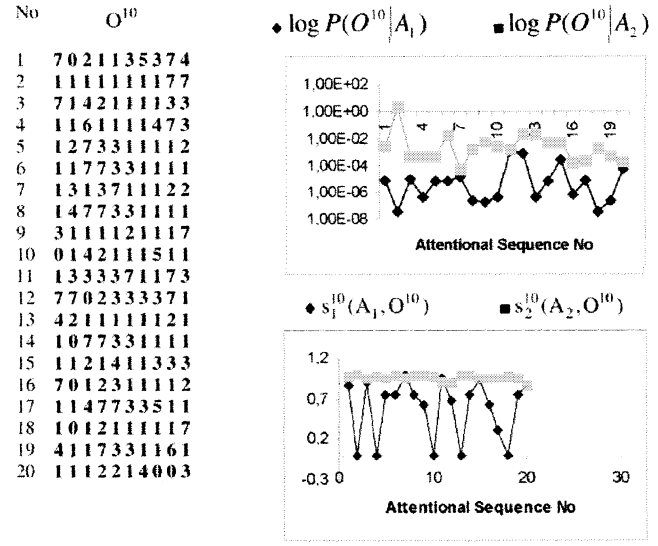


Fig. 10. Results after ten fixations on Scene 2 with ten fixation learning on Scene 1 and Scene 2. Recognition rate is 100% with Markov models and 90% with evidential reasoning.

to that of Scene 2. For recognition experiments, 20 experiments with attentional sequences of length 10 are conducted. Figs. 9 and 10 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 ten fixations during both learning and classification, different feature sequences can be recognized as belonging to the correct shape with a fairly good rate.

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

	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

Fig. 11. Scene 1: 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

Fig. 12. Scene 2: 30 fixation learning.

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 Figs. 11 and 12, this may not be the case. The discriminating transitions 4–6, and 7 between Scene 1 and Scene 3 were better modeled in the previous ten 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.

Results of recognition experiments using models learned from 30 fixations for Scene 1 and Scene 2 are shown in Figs. 13 and 14, respectively. 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 Fig. 14, we had significantly bad results in Experiments 11–14 with both methods. Furthermore, in Fig. 13, where recognition rate was good, both methods returned wrong results in the same two experiments out of 20. The remaining one 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. Figs. 15 and 16 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 Figs. 17 and 18. Results for Object 1 are 100% correct as its model dominates

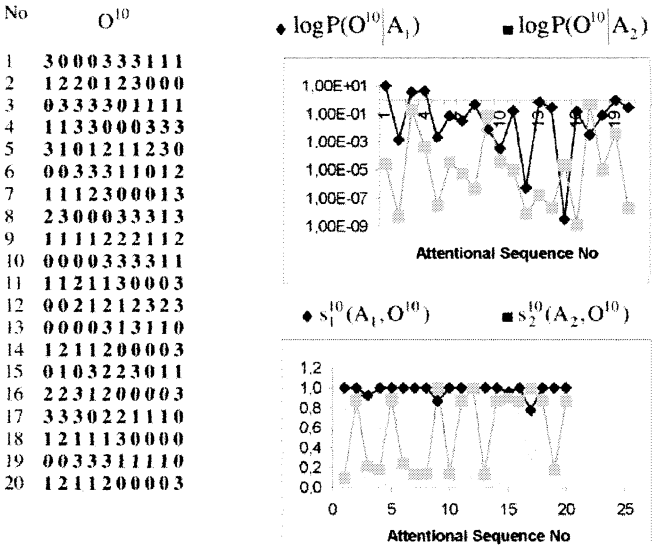


Fig. 13. Results after ten fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 85% with Markov models and 90% with evidential reasoning.

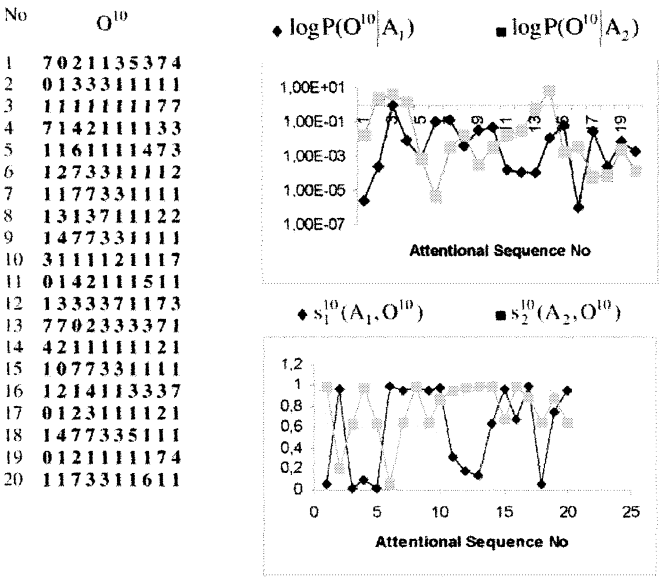


Fig. 14. Results after ten fixations on Scene 2 after 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 50% with Markov models and 60% with evidential reasoning.

	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

Fig. 15. Object 1: 50 fixation learning.

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

	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

Fig. 16. Object 2: 50 fixation learning.

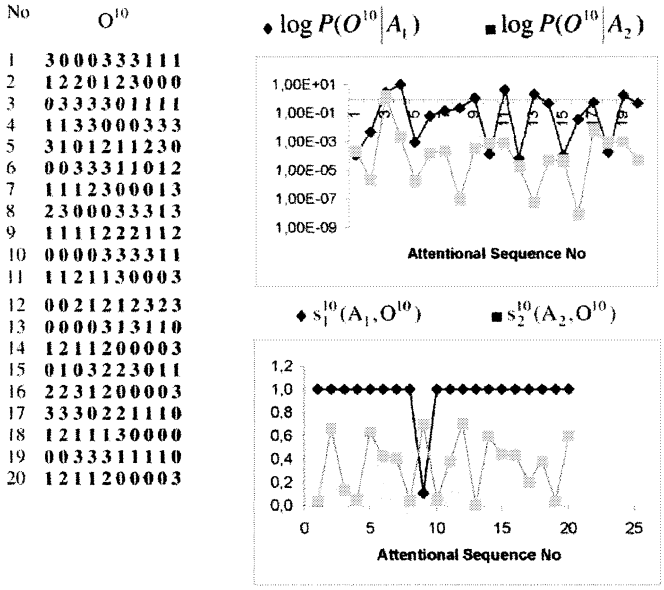


Fig. 17. Results after ten fixations on Scene 1 after 50 fixation learning on Scene 1 and Scene 2. Recognition rate is 85% with Markov models and 95% with evidential reasoning.

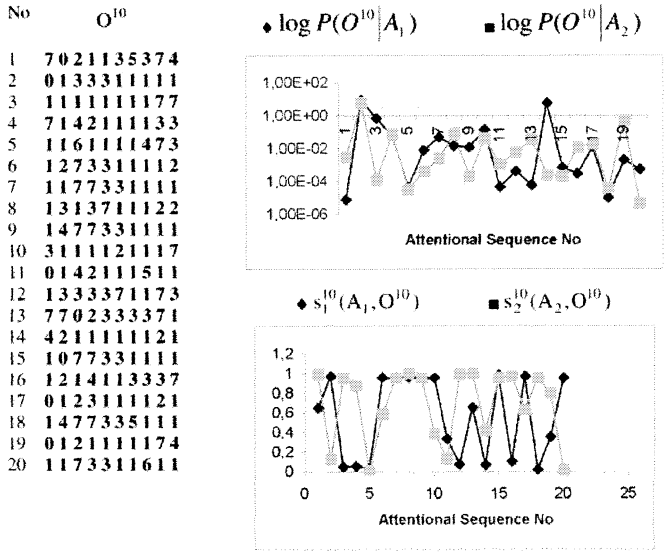


Fig. 18. Results after ten fixations on Scene 2 after 50 fixation learning on Scene 1 and Scene 2. Recognition rate is 50% with Markov models and 60% with evidential reasoning.

very good and in general much better than experiments with ten fixation learning.

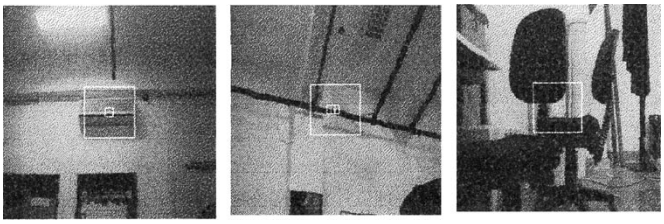


Fig. 19. Wide-angle ten fixations on Scene 1, Scene 2, and Scene 3, as viewed from left to right. Squares represent the visual field and fovea.

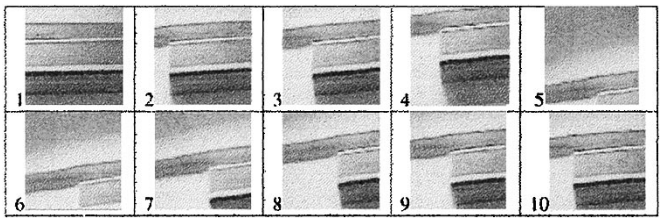


Fig. 20. Sample sequence of visual field images $I_v = (I_v^1, \ldots, I_v^{10})$on Scene 1.

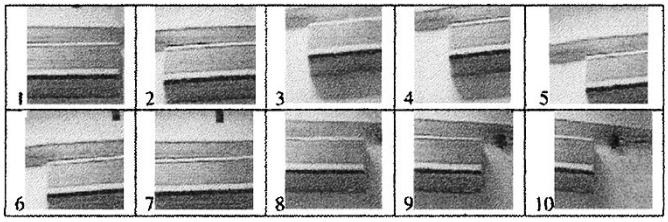


Fig. 21. Second sample sequence of visual field images $I_v = (I_v^1, \ldots, I_v^{10})$on Scene 1.

B. Experiments on Complex Scenes

In the next set of experiments, three complex scenes from our laboratory were used (see Fig. 19). Fixation points and foveas are at the center of each visual field image. Figs. 20 and 21 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, 8–10, 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.

We then compared responses using pairs of models using these complex scenes. Their models were learned using attentional sequences of length 30. Figs. 22, 23–24 give the feature transition frequencies for the three scenes. Simply looking at the generated model matrices, it can be 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

	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

Fig. 22. Scene 1 model.

	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

Fig. 23. Scene 2 model.

	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

Fig. 24. Scene 3 model.

are very similar making classification almost impossible. These results are justified in Figs. 25–30. Scene 3 is recognized with a rate of 100% in all cases as it has a dominating model. In experiments on Scene 1 and Scene 2, results of the two approaches are inconsistent.

C. Experiments on Similar Complex 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 Fig. 31.

In Fig. 32, 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 Fig. 33, 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.

Although these experiments show that scene recognition based on attentional sequences can compensate for small

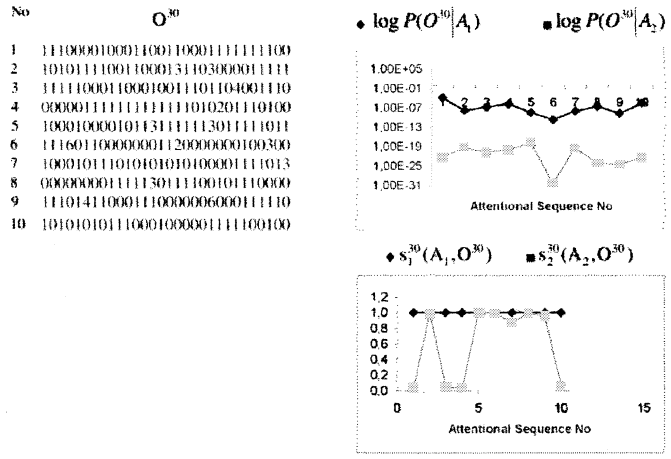


Fig. 25. Results after 30 fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 3. Recognition rate is 100% with Markov models and 100% with evidential reasoning.

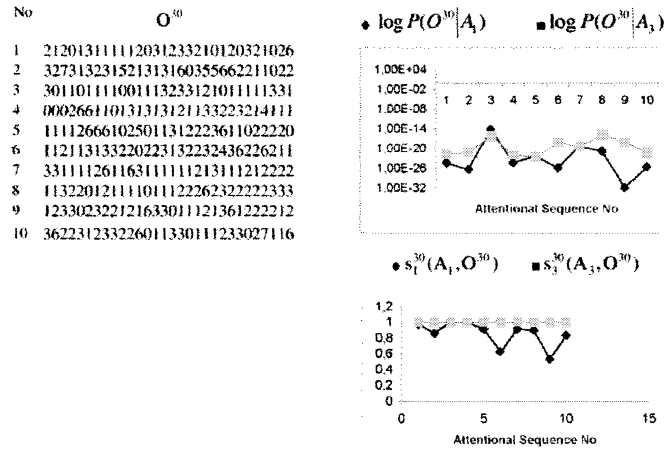


Fig. 26. Results after 30 fixations on Scene 3 with 30 fixation learning on Scene 1 and Scene 3. Recognition rate is 80% with Markov models and 100% with evidential reasoning.

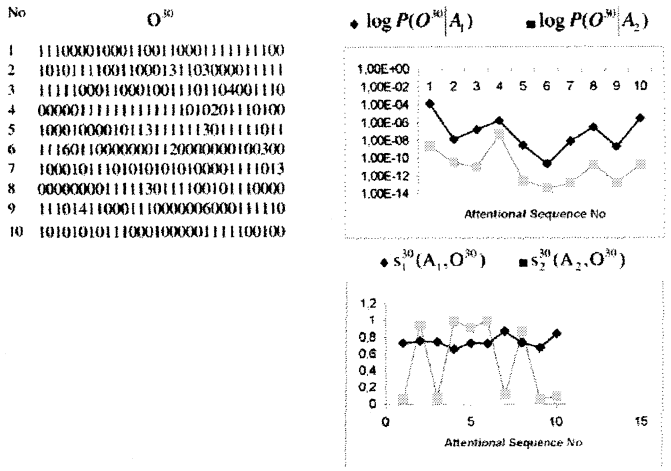


Fig. 27. Results after 30 fixations on Scene 1 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 100% with Markov models and 50% with evidential reasoning.

changes in the environment, the low margins in Scene 2 recognition results in Fig. 32 are confusing. This result may suggest that the model of Scene 1 may be dominating over

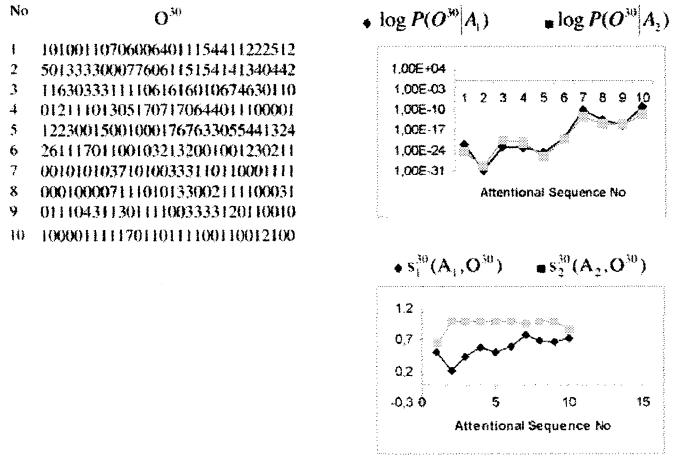


Fig. 28. Results after 30 fixations on Scene 2 with 30 fixation learning on Scene 1 and Scene 2. Recognition rate is 40% with Markov models and 100% with evidential reasoning.

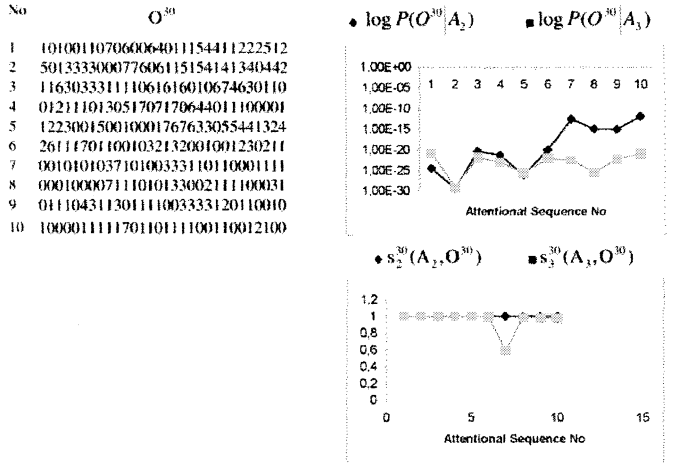


Fig. 29. Results after 30 fixations on Scene 2 with 30 fixation learning on Scene 2 and Scene 3. Recognition rate is 70% with Markov models and 70% with evidential reasoning.

Scene 2 and correct classification of Scene 3 and Scene 4 is a result of this dominance.

D. Summary

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 ten 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.

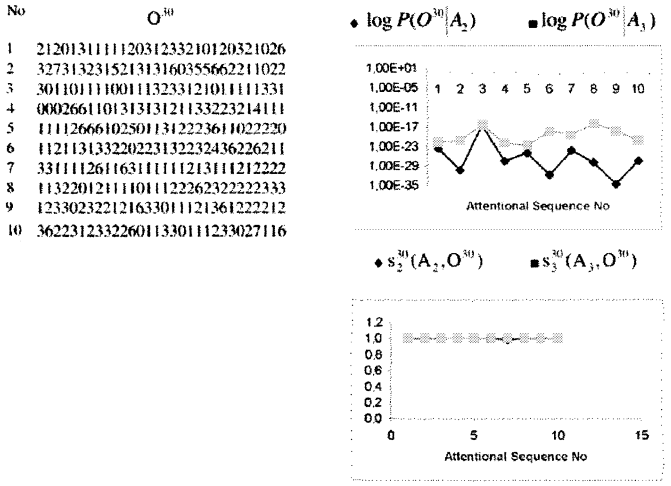


Fig. 30. Results after 30 fixations on Scene 3 with 30 fixation learning on Scene 2 and Scene 3. Recognition rate is 100% with Markov models and 100% with evidential reasoning.

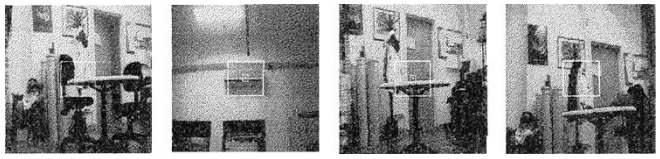


Fig. 31. Wide-angle images of Scene 1, Scene 2, Scene 3, and Scene 4, as viewed from left to right.

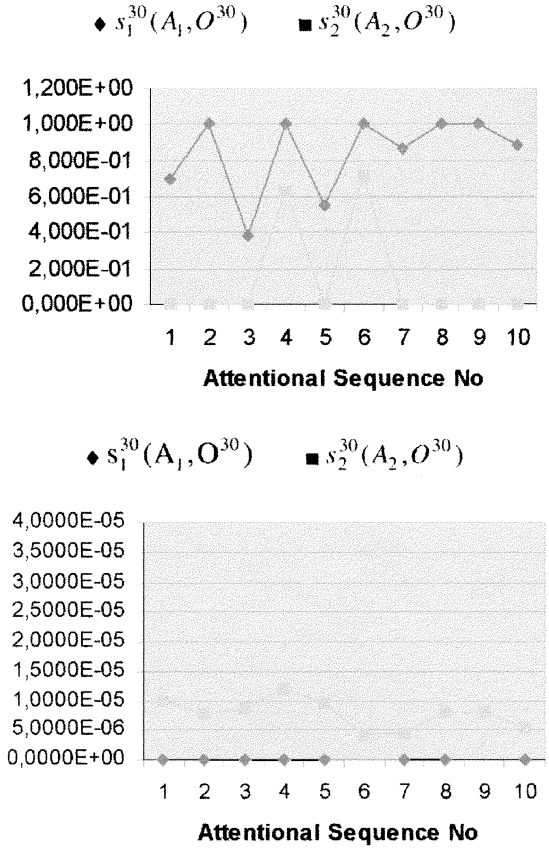


Fig. 32. Results of 30 fixations on Scene 1 (top) and Scene 2 (bottom) after 30 fixation learning on Scene 1 and Scene 2. Recognition rates are 100% and 80%, respectively.

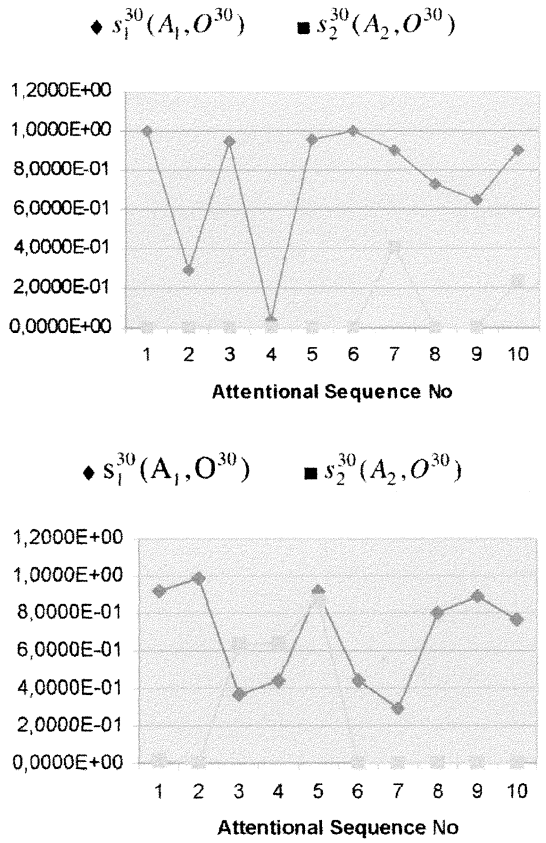


Fig. 33. Results of 30 fixations on Scene 3 (top) and Scene 4 (bottom) after 30 fixation learning on Scene 1 and Scene 2. Recognition rates are 100% and 80%, respectively.

5) Increasing the learning period does not necessarily improve performance. Good performance with a 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 on which we are working.

E. 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, the simple attention function and simple attentive features discussed in Section III 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 [7], [45]. 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 eight 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 IV-C may not be achieved.

As mentioned earlier, 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 subsystems based on the current task specification while keeping the same decision system.

V. Conclusion and Future Work

Biological evidence suggest that, besides being massively parallel, human vision is also sequential especially when solving complex visual tasks. Information is collected in space and time via attention mechanisms resulting in the attentional sequence. In order to better understand human vision and build robots that can parform similarly, we need to learn how to manipulate and use space-time sequences, which are a relatively new data type for vision scientists. In this paper, we propose two approaches to using attentional sequences in recognition tasks: 1) Markov and 2) evidential models. Both approaches are implemented and tested on a working active vision system integrated into APES, a mobile robot designed and developed in our laboratory. Experimental results show that both methods can be used as sequence modeling and classification tools in both simple and complex scenes. However, the success of classification is also dependent on the efficiency of learning and the feature space being used. These two determine the information content of library models and observation sequences, respectively.

In our future work, which is inspired largely by work in vision science investigating the orientation, texture, and frequency specific detector cells in the primate visual cortex, APES will use a higher dimensional and more complex feature space and possibly surfaces. This will enable APES to generate much more complicated observation sequence with richer content. In this case, we can also expect good recognition performance in the presence of more than two models. Using different approaches in simulating fovea-periphery distinction is another interesting study that is likely to improve the performance of feature extraction.

REFERENCES

[1] D. H. Ballard, "Animate vision," *Artif. Intell.*, vol. 48, pp. 57–86, 1991.

[2] D. H. Ballard and C. M. Brown, "Principles of animate vision," *CVIP: Image Understanding*, vol. 56, pp. 3–21, July 1992.

[3] A. L. Abbott *et al.*, "Promising directions in active vision," *Int. J. Comput. Vis.*, vol. 11, no. 2, pp. 109–126, 1993.

[4] J. Aloimonos, "Purposive and qualitative active vision," presented at the Image Understanding Workshop, Sept. 1990.

[5] B. Ballard, "On the function of visual representations," in *Perception*, K. Akins, Ed. London, U.K.: Oxford Univ. Press, 1996, pp. 111–131.

[6] H. Barrow and J. M. Tenenbaum, "Computational vision," *Proc. IEEE*, vol. 69, pp. 572–595, May 1981.

[7] H. I. Bozma and Ç. Soyer, "Shape identification using probabilistic models of attentional sequences," presented at the Workshop Machine Vision Applications, Univ. Tokoyo, Japan, 1994.

[8] P. I. Corke, "Visual control of robot manpulators—A review," in *Visual Servoing*, K. Hashimoto, Ed. Singapore: World Scientific, 1993, pp. 1–32.

[9] M. A. Goodale, "The cortical organization of visual perception and visiomotor control," in *Visual Cognition*, S. M. Kosslyn and D. N. Osherson, Eds. Cambridge, MA: MIT Press, 1995, pp. 167–214.

[10] P. Gouras, "Oculomotor system," in *Principles of Neural Science*, J. H. Schwartz and E. R. Kandel, Eds. Amsterdam, The Netherlands: Elsevier, 1986.

[11] P. Gouras and C. H. Bailey, "The retina and phototransduction," in *Principles of Neural Science*, J. H. Schwartz and E. R. Kandel, Eds. Amsterdam, The Netherlands: Elsevier, 1986.

[12] C. Koch and S. Ullman, "Selecting one among the many: A simple network implementing shifts in selective visual attention," MIT Tech. Rep., no. AIM770, Cambridge, MA, Jan. 1994.

[13] C. Koch and L. Itti, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[14] L. Itti and C. Koch, "A comparison of feature-based combination strategies for saliency-based visual attention systems," in *Proc. SPIE Human Vision Electronic Imaging IV*, vol. 3644, San Jose, CA, 1999, pp. 373–382.

[15] A. Treisman and G. Gelade, "A feature integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–136, 1980.

[16] B. Julesz, *Dialogues on Perception*. Cambridge, MA: MIT Press, 1995.

[17] L. Stark and S. R. Ellis, "Scanpaths revisited: Cognitive models direct active looking," in *Eye Movements: Cognition and Visual Perception*, D. F. Fisher, R. A. Marty, and J. W. Senders, Eds. Mahwah, NJ: Lawrence Erlbaum , 1981, pp. 193–226.

[18] M. Henderson, "Visual attention and the attention-action interface," in *Perception*, K. Akins, Ed. London, U.K.: Oxford Univ. Press, 1996, pp. 290–316.

[19] Z. Kapoula and D. A. Robinson, "Saccadic undershoot is not inevitable: Saccades can be accurate," *Vis. Res.*, vol. 26, no. 5, pp. 735–743, 1986.

[20] I. V. Malinov, J. Epelboim, A. N. Herst, and R. M. Steinman, "Characteristics of saccades and vergence in two kinds of sequential looking tasks," *Vis. Res.*, vol. 40, pp. 2083–2090, 2000.

[21] J. Clark, "Spatial attention and latencies in saccadic eye movements," *Vis. Res.*, vol. 39, pp. 585–602, 1999.

[22] K. Nakayama and Z. J. He, "Attention to surfaces: Beyond a cartesian understanding of focal attention," in *Early Vision and Beyond*, T. V. Papathomas, C. Chubb, A. Gorea, and E. Kowler, Eds. Cambridge, MA: MIT Press, 1995, pp. 69–77.

[23] D. Sagi, "The psychophysics of texture segmentation," in *Early Vision and Beyond*, T. V. Papathomas, C. Chubb, A. Gorea, and E. Kowler, Eds. Cambridge, MA: MIT Press, 1995, pp. 69–77.

[24] J. L. Gallant, D. C. Van Essen, and H. C. Nothdurft, "Two-dimensional and three-dimensional texture processing in visual cortex of the Macaque monkey," in *Early Vision and Beyond*, T. V. Papathomas, C. Chubb, A. Gorea, and E. Kowler, Eds. Cambridge, MA: MIT Press, 1995, pp. 89–98.

[25] J. L. Gallant, 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," *J. Neurophys.*, vol. 76, no. 4, pp. 2718–2739, 1996.

[26] C. E. Connor, D. C. Preddie, J. L. Gallant, and D. C. Van Essen, "Spatial attention effects in Macaque area V4," *J. Neurosci.*, vol. 17, no. 9, pp. 3201–3214, 1997.

[27] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. A7, pp. 923–932, 1990.

[28] T. Caelli, "A brief overview of texture processing in machine vision," in *Early Vision and Beyond*, T. V. Papathomas, C. Chubb, A. Gorea, and E. Kowler, Eds. Cambridge, MA: MIT Press, 1995, pp. 89–98.

[29] J. M. Henderson, "Visual attention and the attention-action interface," in *Perception*, K. Akins, Ed. London, U.K.: Oxford Univ. Press, 1996, pp. 290–316.

[30] D. H. Hubel, *Eye, Brain and Vision*. New York: Scientific Amer. , 1988.

[31] J. P. Kelly, "Anatomy of central visual pathways," in *Principles of Neural Science*, J. H. Schwartz and E. R. Kandel, Eds. Amsterdam, The Netherlands: Elsevier, 1986.

[32] E. Kowler, "The role of visual and cognitive processes in the control of eye movements," in *Eye Movements and Their Role in Visual and Cognitive Processes*, E. Kowler, Ed. Amsterdam, The Netherlands: Elsevier, 1990, pp. 1–63.

[33] ——, "Eye movements," in *Visual Cognition*, S. M. Kosslyn and D. N. Osherson, Eds. Cambridge, MA: MIT Press, 1995, pp. 215–266.

[34] D. Marr and E. Hildreth, "Theory of edge detection," in *Proc. Royal Society London*, vol. B207, 1980, pp. 187–217.

[35] J. L. McGaugh, N. M. Weinberger, and G. Lynch, Eds., *Brain and Memory*. London, U.K.: Oxford Univ. Press, 1995.

[36] D. Noton and L. Stark, "Scanpaths in eye movements during pattern recognition," *Science*, vol. 171, pp. 308–311, Jan. 1971.

[37] M. Pavel, "Predictive control of eye movement," in *Eye Movements and Their Role in Visual and Cognitive Processes*, E. Kowler, Ed. Amsterdam, The Netherlands: Elsevier, 1990, pp. 71–112.

[38] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[39] R. D. Rimey and C. M. Brown, "Selective Attention as sequential behavior: modeling eye movements with an augmented hidden Markov model," Dept. Comput. Sci., Univ. Rochester, Rochester, NY, Tech. Rep., Feb. 1990.

[40] R. D. Rimey and C. Brown, "Control of selective perception using Bayes nets and decision theory," *Int. J. Comput. Vis.*, vol. 12, no. 2/3, pp. 173–207, 1994.

[41] A. Rizzi and D. E. Koditschek, "A dynamic sensor for robot juggling," in *Visual Servoing*, K. Hashimoto, Ed. Singapore: World Scientific, 1993, pp. 229–256.

[42] G. Sandini, F. Gandolfo, E. Grosso, and M. Tistarelli, "Vision during action," in *Active Perception*, Y. Aloimonos, Ed. Mahwah, NJ: Lawrence Erlbaum , 1993.

[43] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press, 1976.

[44] Ç. Soyer, H. I. Bozma, and Y. Istefanopulos, "A mobile robot with a biologically motivated active vision system," in *Proc. IEEE RSJ Int. Conf. Intelligent Robots Systems*, 1996, pp. 680–687.

[45] Ç. Soyer and H. I. Bozma, "Further experiments in classification of attentional sequences: Combining instantaneous and temporal evidence," *Proc. IEEE 8th Int. Conf. Advanced Robotics*, pp. 991–996, 1997.

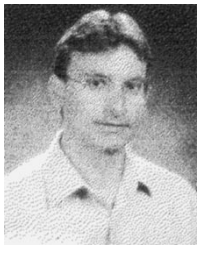
[46] P. Viviani, "Eye movements in visual search: Cognitive, perceptual and motor control aspects," in *Eye Movements and Their Role in Visual and Cognitive Processes*, E. Kowler, Ed. Amsterdam, The Netherlands: Elsevier, 1990, pp. 71–112.

[47] S. Zeki, "The visual image in mind and brain," *Sci. Amer.*, vol. 267, pp. 43–50, Sept. 1992.

[48] C. F. Westin, "Attention control for robot vision," in *Proc. CVPR*, 1996, pp. 726–733.

[49] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L. N. Podladchikova, and N. A. Shevtsova, "A model of attention-guided visual perception and recognition," *Vis. Res., Special Issue Models Recognit.*, 1998.

[50] H. Tagare, K. Toyama, and J. G. Wang, "A maximum-likelihood strategy for directing attention during visual search," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 491–500, May 2001.



Çağatay Soyer received the B.S. degree in electrical and electronic engineering, the M.S. degree and the Ph.D. degree, in biomedical engineering from Boğaziçi University, İstanbul, Turkey in 1993, 1995, and 2002, respectively.

He is currently a Principal Scientist at NATO C3 Agency in The Hague, The Netherlands. Between 1994 and 2001, he was the co-founder of two start-up businesses, Onitron Engineering and VirtualIstanbul.Com., Istanbul, Turkey. Between 1995 and 2002, he was a Lecturer at the Turkish Air Force Academy, İstanbul. His research and professional interests include robot vision, intelligent robots, visual attention, scene recognition, computer graphics, mutimedia, virtual reality, and visual simulation.

H. Işıl Bozma received the B.S. degree (with honors) from Boğaziçi University, İstanbul, Turkey in 1983, the M.S. degrees from Case Western Reserve University, Cleveland, OH in 1985, and Yale University, New Haven, CT in 1986, and the Ph.D. degree from Yale University in 1992, all in electrical engineering.

Since 1992, she has been with Boğaziçi University, where she is currently an Associate Professor in the Department of Electrical and Electronic Engineering. In 1996, she founded the Intelligent Systems Laboratory within the same department and has been the Director since. Since 1994, for various periods, she has been a Visiting Scholar in the Artificial Intelligent Laboratory, University of Michigan, Ann Arbor. Her research interests include intelligent sensor systems, machine vision, robotics, game theory, and automation systems.

Yorgo İstefanopulos (M'78) received the B.S. degree from the American Robert College, Istanbul, Turkey in 1967, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1969 and 1972, respectively, all in electrical engineering.

Since 1972, he has been with Boğaziçi University, Istanbul, Turkey, where he is a Professor in the Department of Electrical and Electronic Engineering. Since 1986, he has been the Chairman of the interdisciplinary graduate program in systems and control engineering and since 1994, he has been Director of the Institute of Biomedical Engineering. His research interests include variable structure control, fuzzy control, intelligent control of rigid and flexible robotic manipulators, system identification, analysis of evoked potentials, spectrum estimation, and signal processing.

Dr. İstefanopulos is the Vice Chairman of the IFAC National Member Organization and he was the General Chair of the 1997 IEEE International Symposium on Intelligent Control and the General Chair of the 2001 Annual International Symposium of the IEEE Engineering in Medicine and Biology Society. He is currently a Partner Member of the Balkan and Eastern European Network of Excellence for the Diffusion of Mathematics for Industry Expertise.