Four Frames Suffice:
A Provisionary Model of Vision and Space

Jerome A. Feldman
Computer Science Department
The University of Rochester
Rochester, NY 14627

TR99
September, 1982

Abstract

This paper presents a general computational treatment of how mammals are able to deal with visual objects and environments. Among the issues addressed are: constancies and the stable visual world, indexing and context effects, perceptual generalization and allocentric spatial maps. The computational model is expressed in connectionist terms, allowing biological as well as psychological experiments to be included. The model relies heavily on contemporary work in Artificial Intelligence, but is claimed to be consistent with all relevant findings.

The preparation of this paper was supported in part by the Defense Advanced Research Projects Agency Grant No. N00014-82-K-0193 and in part by Defense Advanced Research Projects Agency Grant No. N00014-78-C-0164.

1. Introduction

This paper is an attempt to specify a computationally and scientifically plausible model of how mammals perceive objects and deal with their visual environments. The provisionary model is perforce crude, but is claimed to be consistent with all of the known behavioral, structural and computational constraints. The perspective taken is that of a designer of complex information processing systems--one simply sets out to see how a visual system meeting the known behavioral specifications might be built out of the neural componentry, as described in the literature. The resulting four-frames model appears to be a reasonable start.

The rest of this introduction is mainly concerned with describing the main phenomena to be covered by the model and the role of the four representation frames that are the core of the model. The actual specification of the model requires a fair amount of machinery and this is outlined in Section 2. The necessary machinery includes a formal specification of an abstract neural computing unit and a variety of constructions built of these units and their properties. All of this is part of the connectionist modelling (CM) development [Feldman & Ballard 1982; Feldman 1981] and readers familiar with that material will discover nothing new in Section 2.

In Section 3, we describe the four-frames model of vision and space as it would apply to a "Small World" of limited complexity and resolution. By limiting ourselves to six visual features and a 10 x 10 visual map; we are able to describe precisely how the basic operations are intended to work. Section 3 is also oversimplified in that only the main pathways are mentioned and in the suppression of many technical problems in reducing the Small World to the mechanisms of Section 2. Section 3 can be read before Section 2 without much loss, for people who prefer to view the forest before the trees.

The serious work begins in Section 4 where we attempt to carry out the reduction of the four-frames model to CM structures. Although the examples are presented at the scale of the Small World, the computational techniques are claimed to work at realistic scale. The purpose of the section is to confront all the basic computational issues that have come to my attention and to show that none are insurmountable. The solutions are presented at varying levels of detail and some refer to previous computational results. There is no attempt in this section to relate the four-frames model to experimental findings in the behavioral and biological sciences.

Section 5 contains a preliminary attempt to relate the model to experimental findings. The claim that the model is consistent with all established results cannot be tested except by readers such as yourself. What is presented is a range of solidly established findings that fit in well with the current model. Some experiments that could yield challenging results for the current model are also suggested, probably not with sufficient detail.

The discursive comments of Sections 1, 3 and 5 derive from the detailed computational models of Section 4 and may not be easy to interpret in isolation. The particular computational models are intended to show the feasibility of the model and should not be taken too literally. More generally, the provisionary nature of the current model cannot be stressed too strongly. The four frames are an attempt to provide a scaffolding for the establishment of theories of vision and space; if it proves to be useful and none of the scaffolding is visible in the resulting structure, it will have done its work.

The entire development is based on a action-oriented notion of perception. The observer is assumed to be continuously sampling the ambient light for information of current value. We initially consider the issues raised by the four-frames as phenomena to be captured independent of any particular structural model. A "frame" in this view is a set of experiences and experiments that seems to share a common representation. Most people have found the following kind of loose discussion an adequate reason to suppose that we will need at least four frames of reference to describe vision and space.

The representation of information in the first frame is intended to model the view of the world that changes with each eye movement. The second frame must deal with the phenomena surrounding what used to be called "the illusion of a stable visual world." A static observer has the experience of (and can perform as if he held) a much more uniform visual scene than the foveal-periphery first frame is processing at each fixation. One can think of the second frame as associated with the position of the observer's head; this is an oversimplification but conveys the right kind of relation between the first two frames. Of course, neither of these two frames is like a photographic image of the world--as even the most casual examination of the structure of the visual system shows clearly. Light striking the retina is already transformed and the layers of the retina, the thalamus and visual cortex all compute complex functions. The crucial difference between the first two frames is that the first one is totally updated with each saccade and the second frame is not. The current model also assumes that the first (*retinal*) *frame* (RF) computes proximal stimulus features and the second frame captures distal (constancy, intrinsic) features as well as being stable; it is therefore called the *stable feature frame* (SFF). That these two representations of visual information are distinct does not seem an unreasonable hypothesis.

The third and fourth representational frames are both multi-modal and thus unlikely to be the same as the first two. The third representation is not geometrical and will be described in the next paragraph. The fourth, or *environmental frame* (EF), is intended to model an animal's representation of the space around it at a given moment. It captures the information that enables one to locate quickly the source of a stimulus from sound, wind, smell or verbal cue as well as maintaining the relative location of visual phenomena not currently in view. For a variety of reasons, the model proposes a single allocentric environmental frame which gets mapped, by *situation links*, to the current situation and the observer's place in it.

The final representational frame to be considered is the observer's general knowledge of the world, including items not dealing with either vision or space. We follow the conventional wisdom in assuming that this knowledge is captured in abstract or propositional form, modelled in our case by a special kind of semantic network. One kind of knowledge encoded will be the visual appearance of objects. Since the other three representations are geometrically organized, we will refer to the collection of semantic knowledge as the *world knowledge formulary* (WKF), to emphasize its nature as a collection of formulas. The WKF will carry much of the burden for integrating information from the other three frames and is far from adequately worked out in this paper. But all we need for now is the notion that the semantic network representation is likely to be quite different from that of the retinal, stable feature or environmental frame. All of this suggests that even a provisionary model of vision and space will require at least four representational frames; that four frames suffice is the contention of this paper.

The initial exposition of the four frames was based on a static observer and a

basically static environment. Most of the detailed discussions in subsequent sections will retain this restriction, but the model does attempt to cover motion as well. The major additional construct needed for moving objects is to postulate explicitly that the entire system has a second mode of operation, which we call *pursuit mode*. To get a feeling for the difference between the two modes, track your finger as you move it along the second line of text on this page. Now go back and read the line of text, using your finger as a pointer. There is considerable evidence that the pursuit mode is computationally distinct and is used for navigation while moving as well as for tracking. The interactions among the four frames in the model are different in pursuit mode, but we will not discuss these seriously until Sections 4 and 5.

One of the principal devices employed in the current model is the assumption that all the visual features of interest can be reduced to explicit parameter values in some representational space. Typical parameter spaces include color spaces, spatial frequency channels and slant-tilt maps for surface orientation. The mapping of primitive shapes, of textures and of motions to parameter spaces remains problematic, but the model assumes that it must be done. A computational advantage of this total parameterization of visual features is that all the subsequent discussion can be framed as discrete computational problems. More importantly, the assumption that early vision computes discrete values of fixed parameters supports a clear view of phenomena such as apparent motion. From the stream of visual input, the visual system continuously calculates the best fit to the critical parameters. The best fit is, of course, sometimes non-veridical giving rise to apparent motion, shape, etc. If our computational model is sound, then careful study of illusions, meta-contrast, etc., should lead to an understanding of the critical parameters and their possible values. This is the traditional goal of perceptual psychology; an explicit computational model permits the expression of more comprehensive and quantitative theories.

The essential requirement of a computational model of vision and space is that it be massively parallel. In addition to the obvious parallelism of the retina and early vision, we require simultaneous massive interaction between computational units within and across levels of organization of the visual system. By exploiting the reduction of all visual features to explicit parameters we can devote an individual computational unit to each separate value of each parameter and allow all these units to interact. Competing *coalitions* of such units will be the organizing principle behind most of our models. Consider the two alternative readings of the Necker cube shown in Figure 1.1. At each level of visual processing, there are mutually contradictory units representing alternative possibilities. The dashed lines denote the boundaries of coalitions which embody the alternative interpretations of the image. The units connected by circular-tipped arcs are assumed to inhibit one another and the others to excite. The units in Figure 1.1 each represent a distinct entity and are thus like the infamous "grandmother cells." Most of our constructions will employ such dedicated units for simplicity; my suggestions on how this relates to neural encodings are outlined in Section 2 and 5.

Figure 1.1: Necker Cube

The technical tools suggested for describing and analyzing computational systems with billions of interacting units are outlined in Section 2 and are prerequisite for any detailed consideration of the model. For this introductory discussion, we need only keep in mind that all of the computations within and among the four frames are assumed to be continuously interacting across myriad

channels. The need for these multiple interacting computations is most clearly seen in the Stable Feature Frame, the starting point for each of our discussions.

The Stable Feature Frame (SFF) takes its name from its two basic functions in the system. The SFF is intended to be the representation of what was called the illusion of a *stable* visual world. It captures, in a spatially organized buffer, the visual information in the current field of view and is stable over fixation eye movements. The model also suggests that this visual information is held in terms of certain invariant (constancy) *features* of the scene such as size and hue rather than in terms of the immediately sensed values of intensity, retinal projection, etc. The SFF contains a set of spatially registered planes, each of which continuously computes the best value of some constancy feature for every point in the visual field using both retinal input and the current values in all the other planes. The SFF serves partially as a visual buffer memory, but what is stored are features constantly undergoing refinement. It is quite close in spirit to the AI notion of Intrinsic Images [Barrow & Tenenbaum, 1978] as extended by the inclusion of global parameter computations [Ballard, 1981].

The major use of the distal visual feature information captured by the SFF is for *indexing* into models of the visual appearance which are part of one's basic knowledge and thus in the World Knowledge Formulary (WKF). An appearance model is assumed to be a hierarchical structure whose base elements are *visual primitives* each of which can be accessed (indexed) by certain combinations of SFF visual features appearing in the same place. It is obviously easier to match an appearance model to distal features values than to direct image measurements. Recognition of an object or situation is modelled as a mutually reinforcing coalition of active nodes in the WKF. The relaxation of feature and model networks also involves top-down, *context*, links from visual primitives to the feature units that are appropriate. The network representation of a situation includes objects not currently in view and has the links to other modalities.

In my technical sense, a *situation* network in the WKF is a hierarchical structure like a complex object with one additional property. Any WKF situation can become connected by *situation links* from the *Environment Frame* (EF) and thus become the observer's structure for dealing with the space around him at that moment. The Environment Frame is modelled as a tesselation by neural units of the three-dimensional space surrounding the observer. Its mapping to the current WKF situation is allocentric (external) and the changing egocentric position and viewable places are represented by changes in activation of EF units. Moving to a new situation is captured by a discrete switch of situation links, mapping the EF to a different WKF situation network.

The final frame to be outlined here is the first one in the perceptual cycle, the Retinal Frame (RF). The RF is intended to capture all the computational structures which reinitialize with each eye movement. A major problem addressed in the paper is how separate fixations could be integrated effectively. Less attention is given here to the questions of exactly what computations are being carried out for color, texture, motion, etc. because these computer vision questions are under extensive study in our lab [Ballard, 1981] and elsewhere. And, of course, most of the contemporary work in visual system physiology and psychophysics is focused on the retinal frame.

Figure 1.2: Four Frames

The four frames model is mainly an attempt to provide a coherent structure for relating the myriad findings on vision and space. In order to keep the paper of manageable size, emphasis is placed on filling in the gaps between existing theories and models of different aspects of vision and space. Somewhat surprisingly, I have encountered no other contemporary effort to do this, even at a discursive level. There are, of course, a large number of researchers whose ideas have had a marked effect on the enterprise. Barlow's Ferrier Lecture [Barlow, 1981] stresses the use of computational as well as physiological constraints in studying the visual system and suggests an important role for parameter spaces. Among perceptual psychologists, Gregory and Hochberg are closest in spirit to the current enterprise. Haber [Haber, 1982] has recently suggested a synthesis of this line of thought with Gibsonian ideas on early vision and his treatment of low-level vision and space appears to agree with ours.

Our approach to the problem is quite like that of Marr in placing primary emphasis on computational adequacy while requiring consistency with biological and behavioral findings. Much of Marr's effort was directed towards problems at a lower level than those addressed here. His primal sketch (augmented with motion, color and disparity data) could serve as our retinal frame. In the areas of overlap, the two models agree on the use of hierarchical, object-oriented descriptions and disagree on the stable feature frame and the importance of context and visual cues other than shape. More generally, our treatment of the SFF and WKF, indexing and context appear to be the natural extension of current Computer Vision practice [Ballard & Brown, 1982], to massively parallel hardware. There has been relatively little computational work on space models [Kuipers, 1973; McDermott, 1980] but what there is fits well into our "situation" treatment. We will discuss how the four-frames model articulates with behavioral and biological studies in Section 5.

The first question one should ask of a model such as the current one is what issues it claims to address. The four-frames model is most concerned with the integration of visual information, and much less with the detailed analysis of color, motion, etc. It purports to say things about eye movements, stability, constancies and how these interact with general world knowledge. Another serious concern is the representation of external space and how this relates to perception and action. All of these considerations are addressed within a computational framework that aspires to be physiologically predictive. The major shortcoming of the current effort, within its own terms, is an inadequate treatment of moving objects and observers. Each of the four frames would require additional machinery to handle movement and changing situations.

Any attempt to describe the phenomena of vision and space must deal with the problems of interactions among the various kinds of representations and computations. Since these interactions are clearly parallel computations in both the channel sense and the multiple-processor sense, a technical discussion will have to use some kind of distributed computation formalism. The particular formalism presented in the next section is adequate to the task and has proved useful in a variety of related problems.

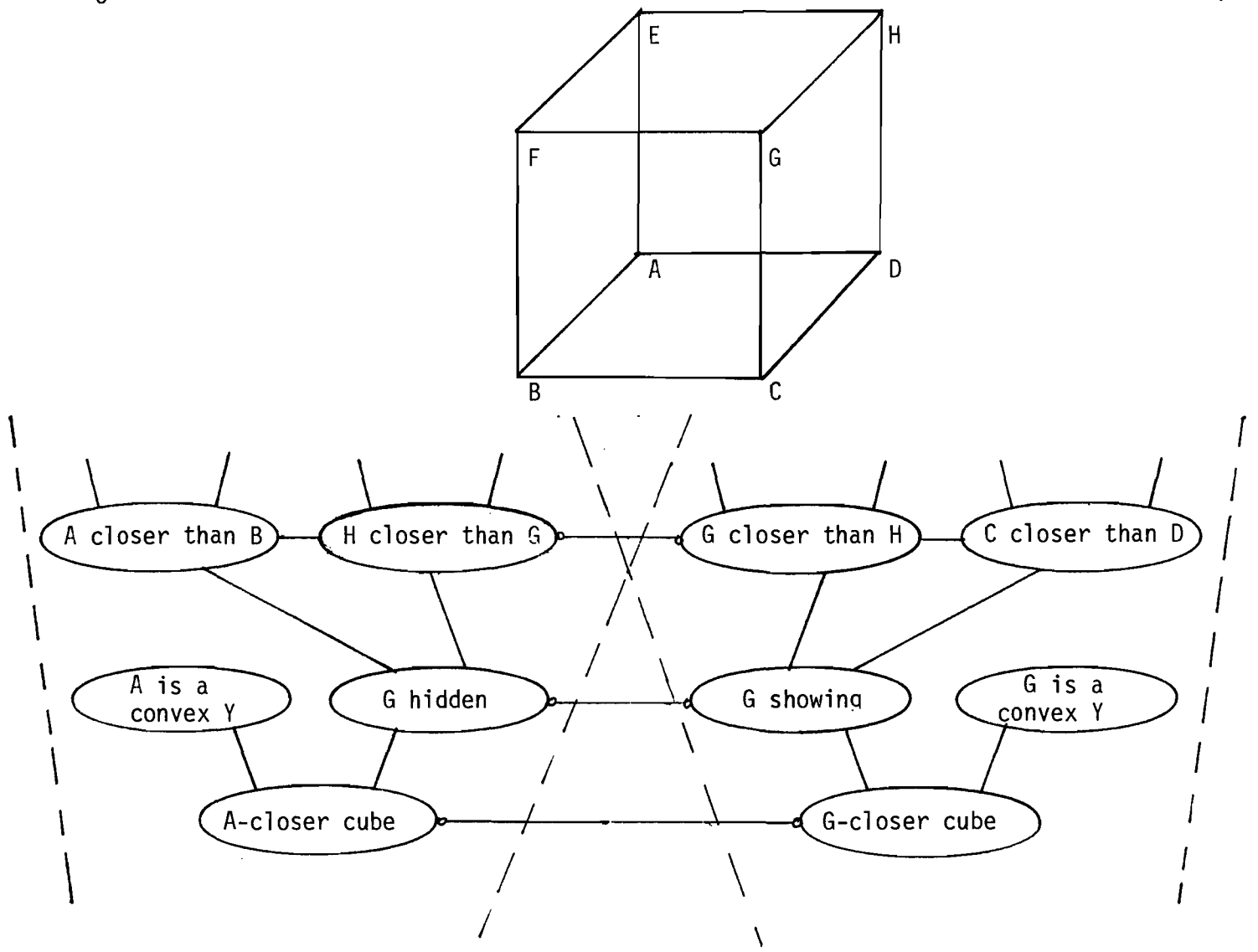


Figure 1.1: The Necker Cube

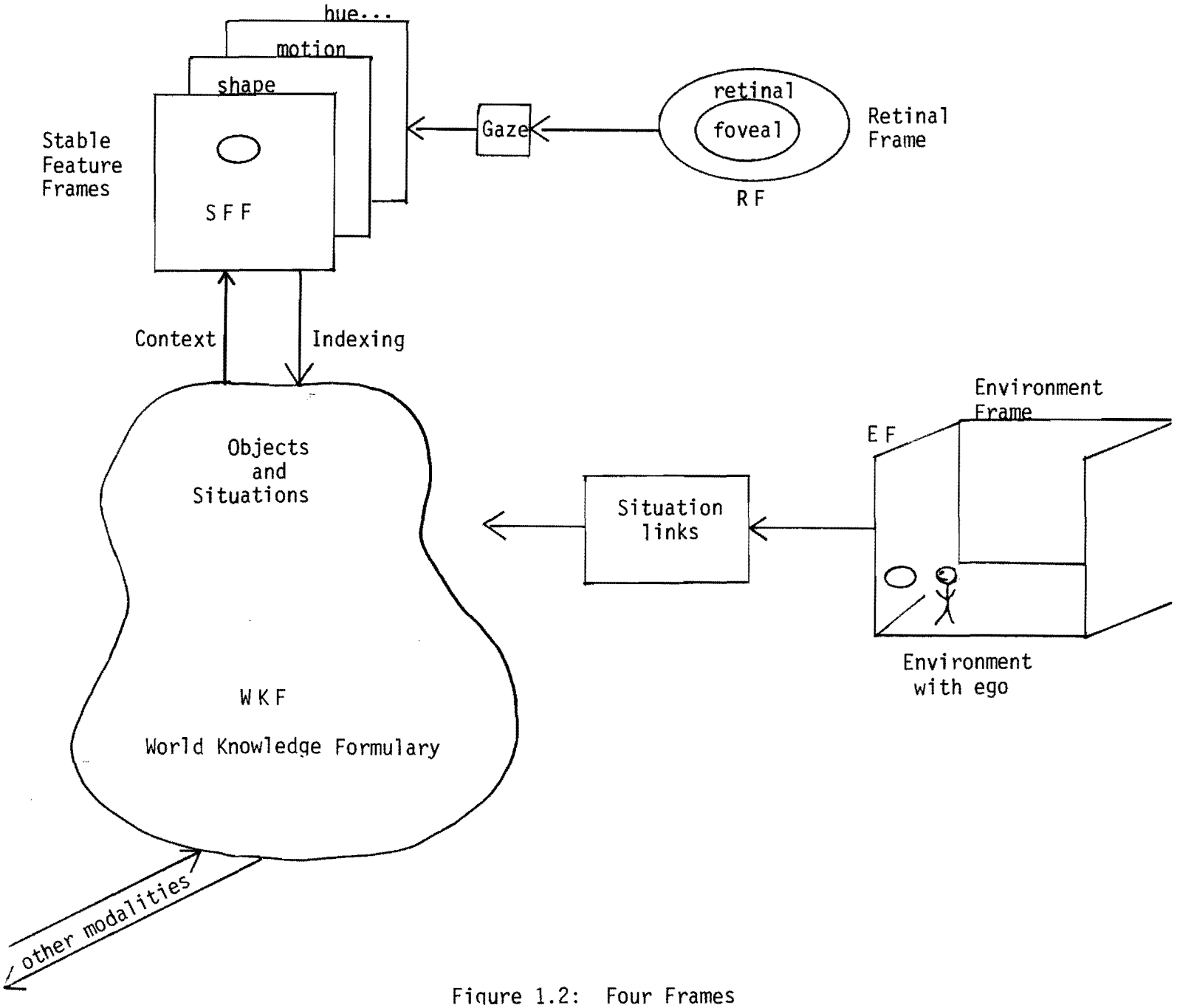


Figure 1.2: Four Frames

2. Connectionist Models

2.1 Background

Much of the progress in the fields constituting cognitive science has been based upon the use of concrete information processing models (IPM), almost exclusively patterned after conventional sequential computers. There are several reasons for trying to extend IPM to cases where the computations are carried out by a massively parallel computational engine with perhaps billions of active units.

Animal brains do not compute like a conventional computer. Comparatively slow (millisecond) neural computing elements with complex, parallel connections form a structure which is dramatically different from a high-speed, predominantly serial machine. Much of current research in the neurosciences is concerned with tracing out these connections and with discovering how they transfer information. Neurons whose basic computational speed is a few milliseconds must be made to account for complex behaviors which are carried out in a few hundred milliseconds [Posner, 1978]. This means that *entire complex behaviors are carried out in about a hundred time steps.* Current AI and simulation programs require millions of time steps.

Various parallel computational models have been successfully used for certain problems in machine perception for some time [Hanson & Riseman, 1978]. What has occurred to us relatively recently is that all of these and more fit nicely into the paradigm of widely interconnected networks of active elements like those envisioned in connectionist models. The generalization of these ideas to the connectionist view of brain and behavior is that all important encodings in the brain are in terms of the relative strengths of synaptic connections. The fundamental premise of connectionism is that individual neurons *do not transmit large amounts of symbolic information.* Instead they compute by being *appropriately connected* to large numbers of similar units. We have been engaged for some time in elucidating the properties of CM models [Feldman & Ballard, 1982; Feldman, 1981] and their application to particular problems in vision reseach [Ballard, 1981]. This paper is the first of this series to attempt a general description of a major function--the perception of objects in space. The plan is to continue to address hard problems (e.g. ambiguity in natural language [Small, 1982]) in technical CM terms so long as it appears to be fruitful.

2.2 Units

As part of our effort to develop a generally useful framework for connectionist theories, we have developed a standard model of the individual unit. It will turn out that a "unit" may be used to model anything from a small part of a neuron to the external functionality of a major subsystem. But the basic notion of unit is meant to loosely correspond to an information processing model of our current understanding of neurons.

Our unit is rather more general than previous proposals and is intended to capture the current understanding of the information processing capabilities of neurons. Among the key ideas are local memory, non-homogeneous and non-linear functions, and the notions of mutual inhibition and stable coalitions.

A **unit** is a computational entity comprising

{q} -- a set id *discrete states*, < 10
p -- a continuous value in [-10,10], called *potential* (accuracy of several digits)
v -- an *output value*, integers $0 \leq v \leq 9$
$\underline{i}$ -- a vector of *inputs* $i_1,...,i_n$

and functions from old to new values of these

$p \leftarrow f(\underline{i},p,q)$
$q \leftarrow g(\underline{i},p,q)$
$v \leftarrow h(\underline{i},p,q)$.

The form of the f,g, and h functions will vary, but will generally be restricted to conditionals and simple functions. Most often, the potential and output of a unit will be encoding its *confidence*, and we will sometimes use this term. The "←" notation is borrowed from the assignment statement of programming languages. This notation covers both continuous and discrete time formulations and allows us to talk about some issues without any explicit mention of time.

The restriction that output take on small integer values is central to our enterprise. The firing frequencies of neurons range from a few to a few hundred impulses per second. In the 1/10 second needed for the basic mental events, there can only be a limited amount of information encoded in frequencies. The ten output values are an attempt to capture this idea.

The inclusion of a discrete set {q} of different states has both biological and computational advantages. It allows the system to accommodate models of fatigue, peptide modulators and other qualitative state changes. Computationally it permits the use of analysis and proof techniques from computer science.

For some applications, we will be able to use a particularly simple kind of unit (p-unit) whose output v is proportional to its potential p (rounded) when $p \geq 0$ and which has only one state. In other words

$p \leftarrow p + \beta \Sigma w_k i_k$ $\qquad [0 \leq w_k \leq 1]$
$v \leftarrow$ if $v > 0$ then round $(p - \theta)$ else 0 $\qquad [v = 0...9]$

where β, θ are constants and w_k are weights on the input values. The weights are the sole locus of change with experience in the current model. The p-unit is somewhat like classical linear threshold elements (Perceptrons [Minsky and Papert, 1972]), but there are several differences. The potential, p, is a crude form of memory and is an abstraction of the instantaneous membrane potential that characterizes neurons; it greatly reduces the noise sensitivity of our networks.

A problem with the definition above of a p-unit is that its potential does not decay in the absence of input. This decay is both a physical property of neurons and an important computational feature for our highly parallel models. One computational trick to solve this is to have an inhibitory connection from the unit back to itself. Informally, we identify the negative self feedback with an exponential decay in potential which is mathematically equivalent. With this addition, p-units can be used for many CM tasks of intermediate difficulty. The Interactive Activation

models of [McClelland & Rumelhart, 1982] can be described naturally with p-units, and some of our own work [Ballard, 1981] and that of others [Marr & Poggio, 1976] can be done with p-units. But there are a number of additional features which we have found valuable in more complex modeling tasks [Feldman & Ballard, 1982].

It is both computationally efficient and biologically realistic to allow a unit to respond to one of a number of alternative conditions. In terms of our formalism, this could be described in a variety of ways. One of the simplest is to define the potential in terms of the maximum of the separate computations, e.g.,

$$p \leftarrow p + \beta \mathrm{Max}(i_1+i_2-\varphi,\ i_3+i_4-\varphi,\ i_5+i_6-i_7-\varphi)$$

where β is a scale constant as in the p-unit and φ is a constant chosen (usually > 10) to suppress noise and require the presence of multiple active inputs [Sabbah, 1981].The max-of-sum unit is the continuous analog of a logical OR-of-AND (disjunctive normal form) unit and we will sometimes use the latter as an approximate version of the former. The OR-of-AND unit corresponding to the definition above is:

$$p \leftarrow p + \alpha \text{ OR } (i_1 \& i_2,\ i_3 \& i_4,\ i_5 \& i_6 \& (\text{not } i_7))$$

Most of the constructions in later sections will employ these "conjunctive connection" units.

2.3 Networks of Units

A very general problem that arises in any distributed computing situation is how to get the entire system to make a decision (or perform a coherent action, etc.). One way to deal with the issue of coherent decisions in a connectionist framework is to introduce winner-take-all (WTA) networks, which have the property that only the unit with the highest potential (among a set of contenders) will have output above zero after some settling time (Fig. 2.1). There are a number of ways to construct WTA networks from the units described above, and several of these have been disccussed in [Feldman & Ballard, 1982] and elsewhere. For our purposes it is enough to consider one example of a WTA network which will operate in one time step for a set of contenders each of whom can read the potential of all of the others. Each unit in the network computes its new potential according to the rule:

$$p \leftarrow \text{if } p > \max(i_j, .1) \text{ then } p \text{ else } 0.$$

Figure 2.1: Winner-Take-All network.

A problem with previous neural modeling attempts is that the circuits proposed were often unnaturally delicate (unstable). Small changes in parameter values would cause the networks to oscillate or converge to incorrect answers. What appears to be required are some building blocks and combination rules that preserve the desired properties. For example, the WTA subnetworks of the last example will not oscillate in the absence of oscillating inputs. This is also true of any symmetric mutually inhibitory subnetwork.

Another useful principle is the employment of lower-bound and upper-bound cells to keep the total activity of a network within bounds. Suppose that we add two extra units, LB and UB, to a network which has coordinated output. The LB cell compares the total (sum) activity of the units of the network with a lower bound and sends positive activation uniformly to all members if the sum is too low. The UB cell

inhibits all units equally if the sum of activity is too high. Under a wide range of conditions (but not all), the LB-UB augmented network can be designed to preserve order relationships among the outputs v_j of the original network while keeping the sum between LB and UB. We will often assume that LB-UB pairs are used to keep the sum of outputs from a network within a given range. This same mechanism also goes far towards eliminating the twin perils of uniform saturation and uniform silence which can easily arise in mutual inhibition networks. Thus we will often be able to reason about the computation of a network assuming that it stays active and bounded.

For a massively parallel system such as the ones we are envisioning to make a decision (or do something), there will have to be states in which some activity strongly dominates. One example of this is the WTA network. But the general idea is that a very large complex subsystem must stabilize, e.g. to a fixed interpretation of visual input. The way we believe this to happen is through mutually reinforcing coalitions which dominate all rival activity for a period of time. Formally, **a coalition will be called stable when the output of all of its members is non-decreasing.** Notice that a coalition is not a particular anatomical structure, but a temporarily mutually reinforcing set of units, in the spirit of Hebb's cell assemblies [Jusczyk, 1980].

The mathematical analysis of CM networks and stable coalitions continues to be a problem of interest. We have achieved some understanding of special cases [Feldman & Ballard, 1982] and these results have been useful in designing CM too complex to analyze in closed form [Sabbah, 1981].

By combining the ideas of conjunctive connections, WTA and stable coalitions, we can developnetworks of considerable power and flexibility. Consider the example of the relation between depth, physical size, and retinal size of a circle. (Assume that the circle is centered on and orthogonal to the line of sight, that the focus is fixed, etc.) Then there is a fixed relation between the size of retinal image and the size of the physical circle for any given depth. That is, each depth specifies a *mapping* from retinal to physical size (see Fig. 2.2).

Figure 2.2: Relations among depth, retinal size, and physical size.

Here we suppose the scales for depth and the two sizes are chosen so that unit depth means the same numerical size. If we knew the depth of the object (by touch, context, or magic) we would know its physical size. For example, physical size = 4 and depth = 1 make a *conjunctive connection* with retinal size = 4. Each of the variables may also form a separate WTA netowrk; hence rivalry for different depth values can be settled via inhibitory connections in the depth network. Notice that this network implements a function phys = f(ret,dep) that maps from retinal size and depth to physical size, providing an example of how to replace functions with parameters. For the simple case of looking at one object perpendicular to the line of sight, there will be one consistent *coalition* of units which will be *stable*. The network does something more; the network can represent the consistency relation R among the three quantities: depth, retinal size, and physical size. It embodies not only the function f, but its two inverse functions as well (dep = f_1(ret,phys), and ret = f_2(phys,dep)). Much of the vision work in our lab [Ballard, 1981] and elsewhere [Hanson & Riseman, 1978] relies on the interaction among constraint networks like those of Figure 2.2.

The stable coalition mechanism also has implications for the "grandmother cell" issue. Even the 3-unit loop capturing a size-depth relationship could be viewed as a "pattern of activity" of the three units. More generally, in any CM network, there will always be many active units forming one or more coalitions. This does not mean that one can usefully characterize the network in terms of diffuse system states instead of units with particular functions. On the other hand, a unit will participate in several coalitions and need not have a simple response pattern. There are both biological and computational advantages to employing the simultaneous activity of multiple units to code some information of interest.

For example, suppose we wanted to represent 10 values each of ten low-level visual features such as position, orientation, hue, contrast, motion, etc. Having a separate unit for each vector of values would require 10^{10} units which is clearly too many. Suppose instead we had units which were precise in only one dimension. Then we would need only 10 x 10 units but it would take the simultaneous activity of ten units to specify a full vector of values. There are a range of intermediate constructions [Hinton, 1981; Feldman & Ballard, 1982]. One of these techniques (coarse-fine tuning) appears close to the coding used in primary visual cortex, where units are broadly tuned in several dimensions and fine-tuned in one stimulus dimension. Consideration of the particular coding techniques employed by the brain will be deferred until Section 5 and we will use whatever coding seems easiest to understand in earlier sections.

2.4 Memory and Change

In the previous section, we saw how fixed CM networks could be designed to compute functions and relations quite efficiently. These fixed networks could have a certain amount of built-in flexibility by explicitly incorporating *parameters*. One can view the depth networks of Figure 2.2 as computing the physical size of objects from the retinal size, parameterized by depth.

But there are also a number of situations where it does not seem plausible to assume the existence of either fixed or parameterized links. An obvious, though artificial, set of examples are the paired-associate tasks with nonsense syllables used by psychologists. A closely related real task is learning someone's name or the Hebrew word for apple. One cannot assume that all the required connections are pre-established, and it is known that they do not grow rapidly enough (in fact, very little at all) [Cotman, 1978]. What does seem plausible is that there is a built-in network, something like a telephone switching network, which can be configured to capture the required link between two units. We refer to this as establishing a "dynamic connection" in the uniform network. We are assuming (as is commonly done) that the weight of synaptic connections cannot change rapidly enough to do this, so that all dynamic connections are based on changes in the potential (p) and state (q) of individual units. The other basic constraints that we impose on possible solutions are that units broadcast their outputs and that there is no central controller available to set up the dynamic connections. These assumptions differ from those in the switching literature, and the results there don't carry over in any obvious way. The assumption is that only one dynamic connection is made at a time, but that several (e.g. 7 ± 2) must be sustainable without cross talk.

The example task we will be considering is to make arbitrary dynamic connections between two sets of units labelled A. . .Z and a. . .z respectively. These could be words in different languages, paired associates, words and images, and so

on. Figure 2.3 depicts the situation for three units on each side.

The problem is how to establish, for example, the link B-c without also linking, e.g. B-b, since the network is originally uniform. More precisely, we require an algorithm which, given the simultaneous activation of B and c, will establish p and q values in the units of our network such that (for some time) activating B will stimulate c but not a or b. For the most part we will consider symmetric networks where the "dynamic connection" B-c will also have the activation of c stimulate B and not A or C. It should be clear that primitive units without any internal state (memory) will not be usable in such tasks.

The basic solution to the dynamic link problem in CM networks relies upon mutual inhibition between the alternative inter-units. For notational convenience, we will sometimes represent this situation as an array of units, with the understanding that the array is a winner-take-all (WTA) network. If the only active link were B-c, then only the three starred units would be active.

Figure 2.3: Uniform dynamic link network.

The idea here is that there is a separate intermediate unit dedicated to each possible pairing. The starred unit for B-c is in two WTA networks, the column which is "inputs to c", and the "outputs from B" WTA net which is drawn in explicitly. When B-c is active, it blocks all others uses of both B and c, which is the desired effect. The fact that our solution requires N^2 intermediate nodes to connect 2N units makes it impractical for linking up sets of 10^5 units like an educated person's vocabulary. There are, however, more complex interconnection networks which require about $4N^{3/2}$ units [Feldman, 1981]. This paper also gives detailed descriptions of the unit computations required and some examples.

2.5 Random Interconnection Networks

There are both anatomical [Buser, 1978] and computational reasons for looking carefully at random interconnection schemes. We will first consider the possibility of using random interconnection networks (in place of the uniform networks above) to dynamically connect arbitrary pairs of units from two distinct layers. As before, each unit is postulated to have links to some large number of intermediate units, whose role is strictly a linking one. In any random connection scheme there will be some finite probability that the required path is simply not present. The remarkable fact is that this failure probability can be made vanishingly small for networks of quite moderate size [Feldman, 1981]. The idea is to have k (two or more) layers of intermediate units so that there is a tree of B^{k+1} links across the network, where B is the outgoing number of branches from each unit. This result has been known for some time and has been used as the basis of a proposed highly parallel computer [Fahlman, 1980].

It is premature to speculate on the degree to which animals are more like the uniform or random networks (if either) but we can say something about the computational advantages of each. Uniform networks appear to be most useful for maintaining many simultaneous dynamic links which are easily turned on and off. They could only be expected to occur in well-structured stable domains because of the strong consistency requirements. In general, we would like to view uniform dynamic links as a mechanism roughly equivalent to modifiable or conjunctive

connections where the number of possibilities is too great to wire up directly.

Random interconnection networks are not as stable and predictable as uniform ones, but have some other advantages. The lower requirements on the number and precision of wiring of intermediate units are clearly important. But the most interesting property of the random networks is the relative ease with which they could be made permanent. Suppose that instead of rapid change we wanted relatively long term linkage of units from the two layers. Our model specifies that this must be done by changing connection weights w_j. The point to be made here is that the random networks already have some units biased towards linking any particular pair from the two layers. By selectively strenghthening the active inputs (on command) of the most appropriate units, the network can relatively quickly forge a reliable link between the pair. The details of how we propose that this comes about are given in [Feldman, 1981] and summarizxed in Section 2.5. Of course, one this has happened, the network will not be able to represent competing dynamic links, but its ability to capture new pairings will remain intact until a large fraction of the nodes are used up (cf. [Fahlman, 1980]).

The fact that random (as opposed to uniform) interconnection networks could be readily specialized suggests that random networks may play an important role in permanent change and memory. After enough training, the originally random interconnection network would become one in which there was essentially a hard-wired connection between particular pairs of units from the two spaces.

The problem with this scheme as a proto-model of long term memory is that most of our knowledge is structured much more richly than paired associates. It is technically true that one can reduce any relational structure to one involving only pairings, and Fahlman [1980] suggests that the best current hardware approach is along these lines. But the intuitive, psychological and physiological [Wickelgren, 1979] notions of conceptual structures involve the direct use of more complex connection patterns. It turns out that the results of the previous section on random interconnection layers extend nicely to the more general case.

The basic situation is shown in Figure 2.4. There are again N (= 16) units connected to $\sqrt{N}$ others, but without any layer structure. We are assuming that all units and connections are identical and that each unit has, at each time step

$$v \leftarrow 2p$$
$$p \leftarrow p + \Sigma i - 2 \ (= \text{decay when } p \neq 0).$$

We suppose that at each time step the unit subtracts 2 from its current potential if not zero, and then adds the sum of its input values. The table in Figure 2.4 shows successive values of p for various units, assuming that at T = 0, units F and I have p = 10 and are maintained for six time steps. The unit O happens to be directly connected to F and I and thus will eventually saturate (under the rules above).

Figure 2.4: Random chunking network.

After step 5, the coalition (F,O,I) is self-sustaining and would actually need to be stopped by fatigue or an external input. In some sense, we can view this coalition as having **recruited** unit O to maintain the dynamic link between F and I. The main differences from the examples given earlier is that here the linking can take place between any set of units and there is no distinction between end and intermediate

units. This is a simple example of the basic mechanism which we believe to support associative learning and appears to be close to what Wickelgren [1979] had in mind. If random chunking networks can be made to support short-term associations through coalitions, the usual weight-changing algorithms would enable the associations to be made permanent.

2.6 Changing Weights and Long-Term Memory

There was a brief discussion of changing weights earlier where it was suggested that random networks could easily be made to incorporate long-term change. We will examine this problem more carefully in this section, still within the constraint that all long-term change is caused by structural modification of connection weights, w_j. There is some evidence for the growth of new connections in adults [Buser, 1978], and for relatively rapid physiological change at synapses [Kandel, 1976], but neither seems to be nearly widespread or selective enough to play a dominant role in the acquisition of knowledge. The discussion in this section will be mainly technical, dealing with rules for changing weights, their properties, and some basic problems.

The standard basis of weight-changing algorithms [Sutton & Barto, 1981; Jusczyk & Klein, 1980] is reinforcement of those weights (w_j) whose inputs (i_j) correlate with desired outputs. This is almost trivially correct, but is subject to a wide range of interpretations, some of which won't work. One widely used rule is to always reinforce those w_j for which i_j was active whenever the unit fires (rapidly). This is the rule originally proposed by Hebb [Jusczyk & Klein, 1980] and has been the basis for many studies of plasticity. However, this feedback-free reinforcement rule provides no way for a system to learn from its mistakes and could not be the only rule used in nature.

Our definition of weight changing in the abstract units depends on a hypothesized ability for a unit to "remember" the activity state of its incoming connections for long enough to get feedback. This assumption is commonly made by modelers (e.g., see [Sutton & Barto, 1981]), and has some currency among neurobiologists (e.g., [Stent, 1973]). The idea is that the activity at a receiving site causes chemical changes that persist and remain localized for some time.

The change in weights will be determined by a function of the inputs (i), potential (p), state (q), and outcome value (x) for each unit. The general case includes a provision for dealing with situations where it is not possible to decide immediately whether a given network behavior should be reinforced. We introduce a "memory" vector $\underline{\mu}$ and two functions, c which updates $\underline{\mu}$, and d, which (usually later) uses values of $\underline{\mu}$ to bring about changes in the weights w. The general definitions are given in [Feldman, 1981]. This paper will not deal with deferred outcomes, so that we can use a simplified definition with $\underline{\mu} = \underline{w}$ and c = d. The rule for weight change becomes

$$\underline{w} \leftarrow d(\underline{i},p,q,x,\underline{w}).$$

As an example, let us consider augmenting the random network of Figure 2.4 to enable it to selectively strengthen connections. We will assume that all of the w_j in the network are initially set to .5. The table in Figure 2.4 is still applicable if we assume that all units have output $v = 4p$ (instead of 2p), because the initial weights of .5 will even things out. We will also have to be more precise in our treatment of

bidirectional links. We interpret Figure 2.4 to mean that, for example, unit O has inputs from and (separately) outputs to units F, I, L, and ?. Recruiting units (O, I, F) to form a more permanent chunk would be accomplished by strengthening their mutual positive effects.

The dynamic link established in Figure 2.4 provides the information necessary for a uniform updating algorithm to choose the right weights to change. For example, the system could signal updating weights at time 5 for all units with $p > 8$. The next thing that needs specifying is a particular updating rule. The next thing that needs specifying is a particular updating rule. A typical update rule might be

$$\Delta w_j = \alpha \cdot i_j$$

which increases weights at a rate proportional to the current input level. A well known problem with this rule is that if weights only increase they will often all saturate. One standard solution (e.g., [Sutton & Barto, 1981]), which works well enough in this case, is to have an increase or decrease in weights which depends on the output or potential of the unit. We could do this discretely by setting a conditional $\delta = 1$ if $p > 8$ and $\delta = -1$ if $p \leq 8$. A continuous version could be $\delta = p - 8$, which would greatly penalize active inputs to dormant units. In either case,

$$\Delta w_j = \alpha \cdot i_j \cdot \delta$$

is an acceptable updating rule. Assuming that the fourth input of unit O is idle, the new values of weights on inputs to unit O would be ($\alpha = .1$):

	I	F	L	?
old	.5	.5	.5	.5
continuous	.6	.6	.56	.5
discrete	.55	.55	.53	.5

Notice that the weight on the mystery input remains unchanged because $i_?$ is zero. This might not be desirable if the goal were to cut off other inputs that might cause confusion with the chunk (O, I, F). In general, different structures will be better served by different updating algorithms and one should not expect to find a uniform scheme that will be applicable in all situations. Our major departure from the literature is to allow non-linear updating rules that need not treat all w_j on a given unit identically. This is a natural extension of the more flexible computational rules we have found useful in our detailed models. Many of the results [Sutton & Barto, 1981] on the convergence and stability of correlation weight changing schemes will carry over to rules of our kind. More details on this and related questions can be found in [Feldman, 1981].

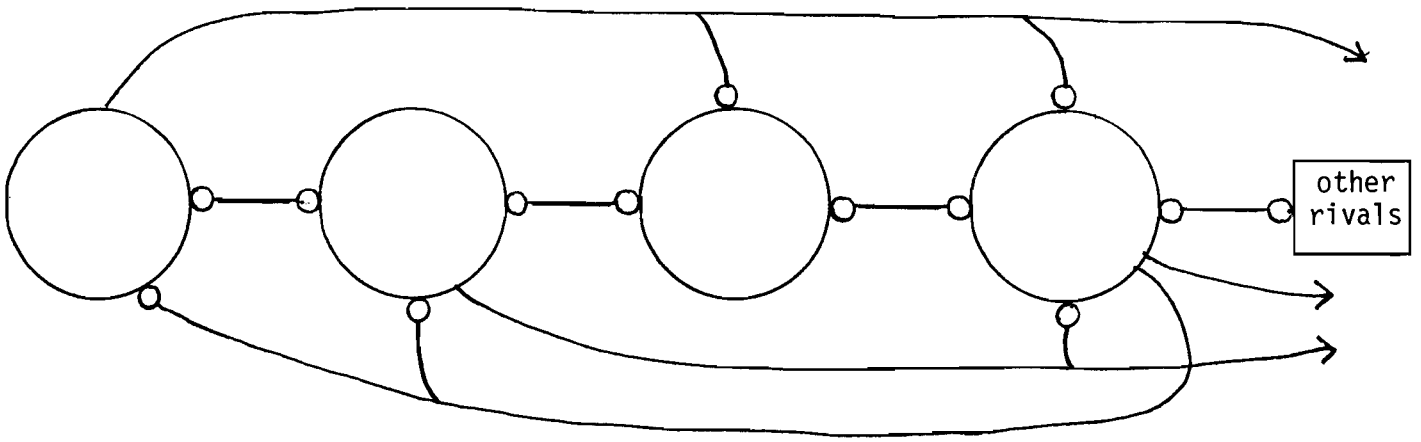


Figure 2.1: Winner-Take-All network. Each unit stops if it sees a higher value.

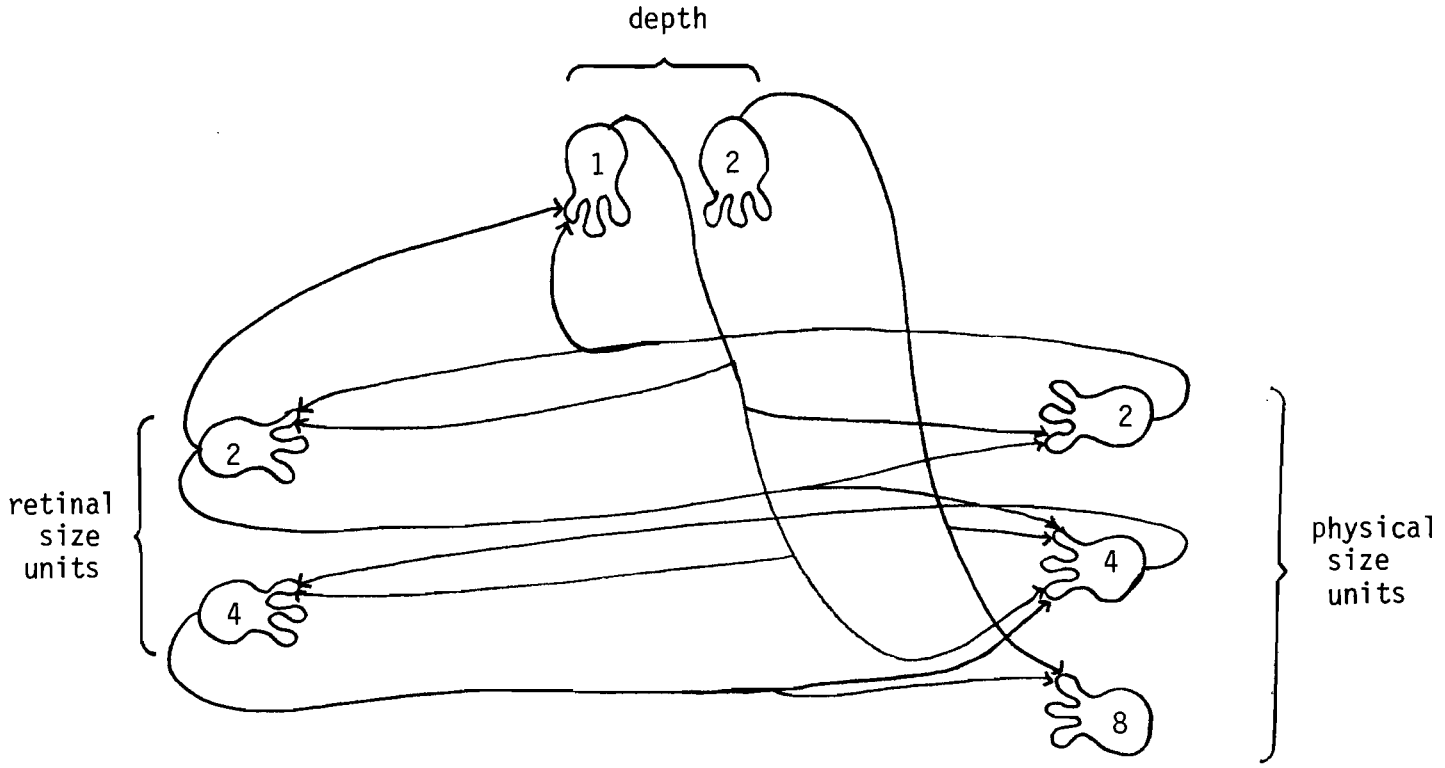


Figure 2.2: Relations among depth, retinal size, and physical size. In the conjunctive depth network, physical size 2 required both retinal size 2 and depth = 1.

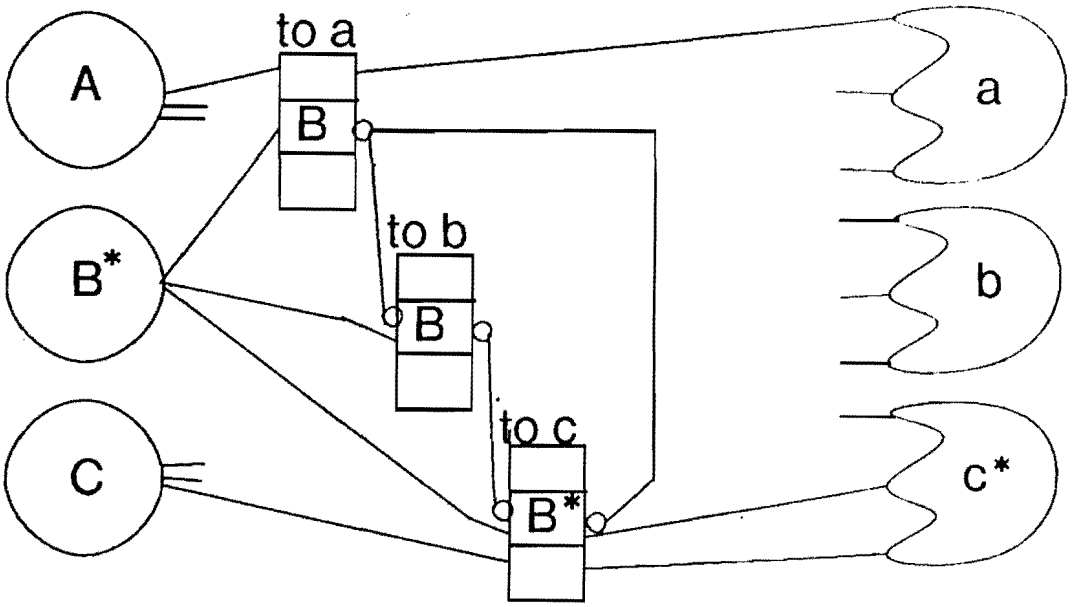
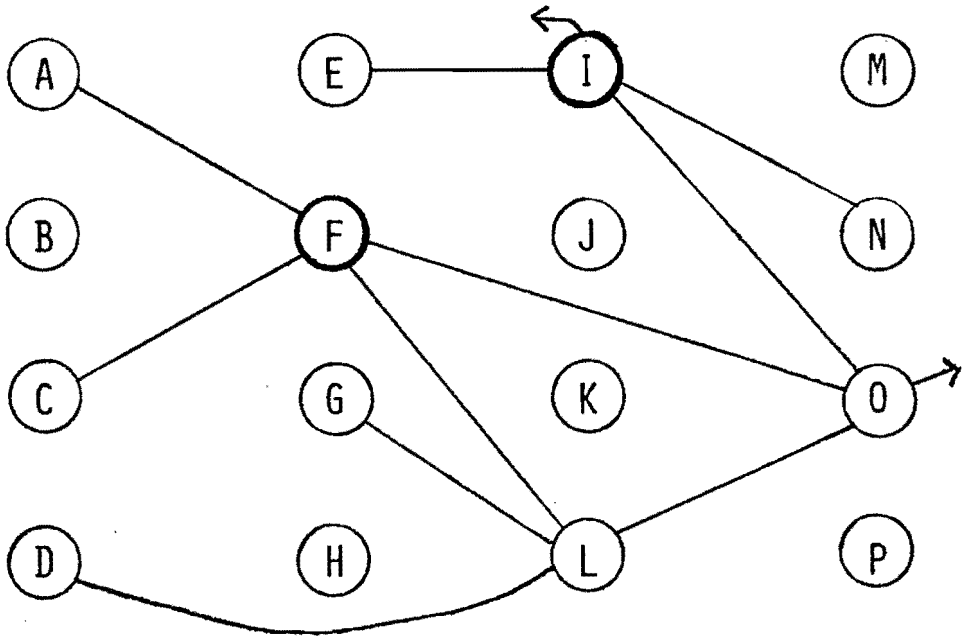


Figure 2.3: Uniform Dynamic Link Network

RANDOM NETWORKS:

N NODES EACH CONNECTED TO $\sqrt{N}$ OTHERS



ASSUME V = .2 * POTENTIAL; DECAY IS 2

	F	I	G	L	O	A	N	. . .
T = 0								
1	10	10	0	0	0	0	0	
2	10	10	0	2	4	2	2	
3	10	10	0	2.8	6	2	2	
4	10	10	1	4	8.6	2	2	
5	10	10	1	6.3	10	2	2	

FIGURE 2.4: RANDOM CHUNKING NETWORK

3. Small World

One problem in trying to think coherently about vision and space is the enormous number of entities involved at every level. In this section we will present a fairly detailed examination of the interactions among the four frames, but all done at a very coarse grain. The small world development has been crucial to the elaboration of the current model and will hopefully also be easier for others to work with. Again, we will push through a straight line of development that ignores many important issues and then try to address all the major ones (in Section 4). This section and the next one still contain no behavioral or physiological support for the choices being made - the concern is strictly with the computational adequacy of the model. Only after the model is specified will we address its relationships with past and future experiments (Section 5).

Our discussion begins with the problem of linking visual feature information with the knowledge of how objects in the world can appear. The problem of going from a set of visual features to the description of a situation will be called the *indexing problem*, following the terminology common in AI. The small world we will consider in detail has exactly six distinct visual features each with 10 possible values (Figure 3.1). Assume for now that any object in the small world can be characterized by some particular set of values for the six features. This would mean that each object has a distinct 6-digit visual code (not unlike a zip code). If the system could always reliably extract the values for the visual features, it would not be hard to identify which objects were in which places in the current environment. No additional problems would arise if some objects had multiple codes among the 10^6 = 1,000,000 available. But the system, as specified, would totally break down if two objects needed to share the same code, i.e. looked identical relative to our set of features and values. We will have to address the question of ambiguous feature sets later.

The six particular visual features which we have chosen are intended to elucidate the major scientific problems in intermediate level vision and would not be the best choice for a practical computer vision system. We assume for now that the best value at each position of the current view is continuously maintained by parameter network computations [Ballard, 1981] which will be elaborated below. Features such as size and shape which cover several units are assumed to be represented by a single unit, say at the center of the region covered. Of course, the problem of breaking up the feature space into meaningful regions is a central one and the model will have to address it in detail.

One of the features which we employ in the small world is called "motion." Motion, as well as the other features, will be treated in this section as a property of objects which has ten discrete values and is continuously updated by computational processes which will be specified later. Motion and change are clearly critical problems and require much more careful treatment than an arbitrary assignment of ten values. But there is an important conceptual advantage to including motion as an explicit parameter even at this early stage. If computing the best discrete valued characterization of object motion is a basic property of low-level vision, then there is nothing at all surprising about the various perceived motion phenomena. More generally, the notion that low-level vision is concerned with continuously maintaining the best current discrete value choices for specific visual features provides a powerful organizing principle for helping to explain a wide range of findings in perceptual psychology. We will consider some of these issues in Section 5,

after the small world model has been worked out in detail.

The model specified so far has almost no content, but several important points can already be seen. The most important point is that discrete values for a fixed set of visual features provide a natural base for indexing, and all of our models will assume this structure. The second point is that the visual features chosen will determine which distinctions the sytem is capable of, as is already well known in classical pattern recognition. An obvious consequence is that the features used for indexing should be as invariant as possible under different viewing conditions. This suggests that we should use the "constancy" properties like reflectance, physical size and surface curvature rather than proximal or image features for indexing.

The six visual features used in indexing are the following: lightness, hue, texture, shape, motion, and size. Obviously enough, ten values of these features (even in logarithmic scales) is not enough to characterize visual appearance in the real world; but the small world is rich enough to exhibit most of the required problems. The model assumes that the six features are continously represented in six parallel 10 x 10 arrays which are intended to map the currently visible external world. There is also assumed to be a (10 valued logarthmic) depth map maintained as part of the same structure (Fig. 3.1). The depth map is needed for calculating constancy features such as object size and is also used directly in mapping the environment. The depth map is assumed to be calculated cooperatively with the six feature planes, using binocular and other cues. These seven parallel arrays, along with some auxiliary structure, comprise the stable feature frame (SFF) which is one of the four cornerstones of the model.

Figure 3.1: The six feature (and depth) planes for the Small World SFF.

The SFF takes its name from its two main properties: it encodes *visual feature values* and it is *stable* over fixations. The SFF is the basic interface between the visual system and the more general world knowledge represented in the World Knowledge Formulary (WKF). The idea is that the SFF at all times maintains a map of the visual properties of the part of the world that is currently in view. We will describe below in some detail how the SFF interacts with the retinal frame (RF) in maintaining a stable visual world. Assuming that the SFF is successfully maintained, we now address the problem of how its feature values can be employed to capture knowledge of the objects in the current environment (and their activities). Thus we return to the indexing problem.

Our first view of appearance models was that each object could be characterized by one or more sets of feature values. For objects that are sufficiently simple, this is not a bad approximation. You can probably name an object that is an approximately 1.5" white sphere and which is uniformly pock-marked even before seeing it hook into the rough. But for complex objects like a horse or Harvard Square, the single feature set isn't even the right kind of visual information. Our way of handling the appearance models for complex objects and situations is, again, taken directly from current AI practice. We assume that the appearance of a complex object is represented (as part of one's world knowledge) as a network of nodes representing the "appearance possibilities" of simpler components and relationships among them. Figure 3.2 shows the description of a chair scene from [Ballard & Brown, 1982] which is typical. There are several unsolved technical questions about the number of separate views maintained, and how much flexibility should be encoded in a description, but the general idea of composition is all we need at the

moment.

Figure 3.2: A typical network representation of visual objects in a situation [Ballard and Brown 1982]

Recall that the naive version of indexing was to use the 6-digit visual feature code to look up the name of the object with that description. Complex objects are assumed to be composed of parts, each part being either another complex object or a *visual primitive* that can be indexed by the 6-digit code. Now recall that all of our structures are assumed to be parallel and continuously active. This means that "indexing" can be continuously in progress between different areas of the SIF and networks of visual appearance knowledge in the WKF. The crude version of this idea is to assume that each set of visual features (for a point in the 10 x 10 SIF map) picks out (indexes) the visual primitive which is appropriate. If this were to happen, it is not hard to see that a visible complex object would have many of its visual primitive parts selected simultaneously and should therefore be recognizable. Parallel indexing from the entire visual field without confusion is too much to expect.

In order to make these notions more precise and eliminate the ghosts from our machine, we must describe all of this in considerably more detail, using the technical definitions of Section 2. The various components of both the SIF and WKF will be elaborated in terms of the "units" of Section 2. Obviously enough, we will need separate units for each of the 100 spatial positions in each of the seven separate maps. In fact, it is also important to follow the unit/value principle and require a separate unit for each value of each cell in the maps above, giving a total of 7000 units. Following the connectionist dogma, we assume that visual primitives are units which are connected to the appropriate set of visual-feature-value units. For example, Figure 3.3 shows how golf and ping pong ball descriptions in the WKF might be connected (indexed) by visual features. It is easy to see how to make connections do the same job as the index codes. Each code for a visual primitive is assumed to be encoded as a conjunction of links from units representing the appropriate value of each feature. A visual primitive with multiple codes has several disjunctive "dendrites," one for each code. Visual primitives that are part of a complex object are also linked into a network for representing the appearance of the object [Figure 3.4].

Figure 3.3: Ping-pong and golf balls

Figure 3.4: Harvard Square situation network
Rectangles are situations, squares are (complex) objects

The general notion of representing a complex object as a network or graph of nodes is standard in machine perception and will be followed here. In the small world we will assume that a node corresponds to one visual primitive (set of feature values) and is represented by a single unit as in Section 2. The links between nodes are assumed to be conceptually labelled as in Figure 3.2. The encoding of labelled links into CM connections will vary, but will mainly be through conjunctive connections involving separate units which embody the link name.

An important aspect of the small world model is that complex *objects* and *situations* have the identical representation as semantic networks in the WKF but

may include several complex objects and relations among them. A *situation* is for us any oriented WKF network which can be mapped to the environmental frame to guide behavior [cf Section 4.2]. The question of whether a given network should be viewed as a situation description is not fixed in advance, but is determined by the way that the description is being used. Intuitively, it seems reasonable enough that a room or Harvard Square can be treated either as a situation or as an object viewed from some distance and that the same relational knowledge could be employed in each use. Both object and situation descriptions allow for nested sub-descriptions and both can accommodate some stylized movement as will be discussed later.

The question of when a network description is playing the role of a situation is quite sharply defined in our model. We assume that at any given time there is exactly one currently active situation description and that it represents the environmental situation at that time. Loosely speaking, the model assumes that there are situation descriptions for places, routes, etc. and that these are linked in the WKF as a "patchwork cognitive map" [Kuipers, 1973]. The technical questions to be addressed here are how these situation descriptions interact with early vision (SFF) and with the (modality-independent) frame which encodes knowledge of the space around us at any time. It is this environmental frame (EF) which is the fourth pillar of our framework; the others being general world knowledge (WKF), features of the stable visual scene (SFF) and the instantaneous retinal information (RF). Again, it is crucial to think of all of these frames as continuously active and interacting with one another.

The environmental frame in the small world is again unrealistically rectilinear. We assume that the world around us is always represented as a box-like three-dimensional spatial map, as shown in Figure 3.5. The nodes of the EF each represent a position in the space surrounding the observer, and the activation of these nodes varies with the direction of gaze. There is a mapping to nodes in the currently active situation (in the WKF) from appropriate units in the environmental frame. Every node in the currently active situation will get some potentiation just from being part of the active situation. Additionally if one of these nodes is mapped to a position in space that is currently being gazed upon, it will receive much more potentiating input and can be said to be "anticipated." Recall that in our discussion of ambiguous visual input we said that mechanisms like this would lead to one interpretation being preferred over another depending on the situation.

Figure 3.5: Two EF units of different scales activate different objects in SFF
435 = Harvard Square

The model includes three levels of top-down input to nodes representing visual objects in the WKF: current situation, visible, and foveated. We will describe the proposed representation for situations and the EF in more detail and worry only later about how one might come to learn the networks for situations (and objects).

Our model of the environmental frame includes a subnetwork for continuously updating the position and orientation of the observer within his environment. This is clearly necessary for computing which parts of the environment are visible and foveated. The same information is assumed to be used in the GAZE mapping linking the retinal and SFF frames. Although it is not so obvious, the ego position within the frame also can provide scale information, allowing us to anticipate more precisely what should be visible from a given view point in the environment. This scale information combines nicely with the hierarchical nature of

the visual descriptions suggested for the WKF. As the observer approaches some object, different levels of substructure become visible and the operation of the current model incorporates this in a natural way. The relative position of objects to the current egocentric position is also assumed to be the basis for physical actions on objects. The model suggests that the SFF-WKF-system is crude and that visual or other sensory guidance is needed for accurate location of objects.

For concreteness, we assume that the (fixed) environment frame has four directions (N,E,S,W); we will not include objects above or below the observer for now. Starting from the center of the map, there are four (logarithmic) distances in each direction. For things at distance one, the observer can resolve 10 x 10 spatial postions. At distance two, the resolution is 5 x 5. At distance 3 it is 2 x 2 and at distance 4 only one unit is active or not. The situations are encoded in a compatible way. Each object description in a situation network has a scale at which it could be visible, if gazed upon.

As the observer moves, the visible scale and postion values are continuously updated. There is no apparent difficulty in also computing occlusion information, either generally through the EF or specifically in the situation description. We assume that situations become mapped as the active current environment, based on how the observer has organized his situation memory. Some general cues as to when situations would change include: going through a door, changing to a different scale of consideration or switching from planning to acting. The technical question is exactly how the environmental frame interacts with the current situation network. The major difficulty is providing for the mapping of a great number of possible situations onto the single fixed environmental frame. Notice that any CM model will face the problem of coupling distrubuted knowledge to fixed input and output systems - the scientific questions are where and how to carry out this coupling. The keys to our solution to the situation - EF mapping problems are: *situation nodes*, conjunctive connections and directly encoding only the inverse mapping. We assume that the environmental frame consists (inter alia) of units that each represent a region of the currently surrounding space. Each of these units will conjunctively connect to all of the objects which might be visible in its region of space in some situation. Not surprisingly, the other half of the conjunctive connection comes from a unit which is active exactly for one particular situation. Figure 3.5 depicts the general situation. If the current situation is "Harvard Square" = S463 then all of the objects in that situation will be receiving some activation. This means that there will be some greater than usual expectation that these objects will be chosen over their rivals in non-visual as well as visual computations. When gaze is of a direction and scale appropriate for some object, its node (in the WKF network) will be more strongly activated because the corresponding position in the EF will be active and this plus the currency of S463 will cause high activation of e.g. "The Coop" and "Brighams". This provides top-down bias to the relaxation between the WKF and the visual features of the SFF, the details of which will be given later. Finally, if a particular known object (say the door of the Coop) is foveated, there will be even stronger top-down bias through the WKF to both the SFF and Retinal computations.

The advantages accruing to a visual system with foveation are the focus of our description of the first basic component of the model - the retinal (RF) frame. Even before we fill in the details we can see that there are several reasons why foveating an object of interest leads to better recognition:

a) Certain complex calculations (e.g. color, texture) can only be done foveally.

b) Bottom-up indexing of features to visual primitives can be restricted to the area of the SFF being foveated (by spatial focus units), greatly reducing the possible confusions.

c) In a known environment, top-down activation from the conjunction of situation and gaze information can significantly raise the activation of an expected object or primitive.

All three of these advantages mutually reinforce one another, leading to an overwhelming advantage for foveal vision in the model. The role of peripheral vision is to set and maintain contexts and to continuously monitor for change, as we will see as the elaboration of the model continues.

The retinal frame (RF) is primarily concerned with bringing the enormous spatial resolution and processing power of the fovea and its maps to bear on points of interest. The RF is assumed to calculate the values of disparity, retinal motion, intensity change, etc. which are the primary inputs to the SFF. The current model assumes that there are local grouping and smoothing processes active within each feature network, but that interactions among features are carried out in the SFF.

In keeping with the rest of the development we will describe a specific incarnation of the retinal frame which is much too small and rectilinear, but should be easier to understand. Our retinal frame will have 100 spatially organized units, like the feature frame (SFF), but they will be laid out very differently. In the RF, 64 of the 100 spatial units will be uniformly packed into an area equivalent to a 2 x 2 array of the SFF. The remaining 36 units will be formed into three surrounding rings of logarithmically decreasing resolution. In terms of SFF units, the units in the outer rings of the retinal frame will cover 1, 4, and 16 squares respectively. All of this is depicted in Figure 3.6.

Figure 3.6: Logarithmic Retinal Frame

We assume that the retinal frame can (logically) move with respect to the SFF. The center of the RF can "move" to any position in the SFF except the two outer most rings. Under these conditions, the entire SFF is covered by the RF at all times. Naturally, the parts of the SFF mapped by the coarse units of the RF get only coarse information while the fovea is mapped elsewhere. Figure 3.7 depicts the situation where the fovea is mapped to the upper left extreme of its range, leaving most of the SFF covered by 2 x 2 and 4 x 4 retinal units.

First, a technical point. The relative motion of the RF must be implemented in our scheme by a switchable conjunctive mapping. We assume that each RF unit is linked appropriately with every combination of SFF units to which it could map. Every such RF-SFF link is conjoined with a connection that specifies the currently active GAZE mapping. For example, in Figure 3.7, the top-left corner unit of the RFF arrays will be mapped to the unit just beyond the fovea which is the top-left of its ring. The mappings for units other than those in this ring are not 1 to 1; this will be important as we consider the interactions of the retinal (RF) and feature (SFF) frames.

Figure 3.7: Retinal Frame mapped to SFF

In the current model, there is no top-down feedback from SFF to RF units. Any tuning of the retinal frame is assumed to be captured in the mechanism for GAZE control. The flow of information in the other direction is, of course, the basic problem of low-level visual processing. The model postulates a distinct fovea and periphery in the retinal frame and assigns quite different functions to them. The fovea (8 x 8 in our case) is assumed to have enough resolution to determine which of the discrete (10 in our case) values of the stimulus features are present in the area foveated. The SFF is assumed to be able to integrate and retain information about hue, texture, shape but not to do the direct computation of the feature values. The main purpose of the SFF is incorporating and maintaining information about the entire visible scene that is only computable foveally. The SFF does not simply transcribe retinal input; the seven planes interact continuously to produce a feature frame which encodes "constancy" values of size, hue, etc. The depth map is needed in the SFF to aide in constancy calculations and, in fact, there appear to be a number of other auxiliary calculations needed as well [Ballard, 1982]. The four units of the SFF currently mapped to the fovea of the RF dominate the calculation of feature values, but an overall consistancy must be maintained.

The peripheral 36 units of the retinal frame are assumed to play a different role. If the SFF is blank, as when a new scene is first encountered, each unit in the RF provides the same value to all the (1, 4, or 16) units in the SFF to which it is currently mapped. These crude values become the basis for the initial relaxation towards constancy features in the SFF and (because they are there) begin indexing the visual primitives in the WKF. This crude indexing is assumed to provide some guidance to the choice of fixation points for further analysis of the scene.

When analysis is well under way and the SFF is not blank, the periphery is assumed to function in a "change detection" mode. The coarse values computed by peripheral units are compared with average values from the (1, 4, or 16) SFF units covered. If there is too large a difference, an alerting signal is activated leading (in the simple case) to a saccade to the place of change. The SFF is also assumed to contain networks for "smooth continuation" of visual properties across fixations. The networks for continuity and "filling in" phenomena are assumed to interact with the coarse values computed by the peripheral RF. There is a wealth of data on visual illusions and meta-contrast phenomena which constrains the choices of how these networks function and interact.

Recall that this entire discussion is ignoring what we have called the "pursuit mode" of the system. In pursuit mode, the periphery does not alert on all changes but is assumed to still be sensitive to optic flow patterns indicating collisions. Pursuit mode is discussed in Section 4.4.

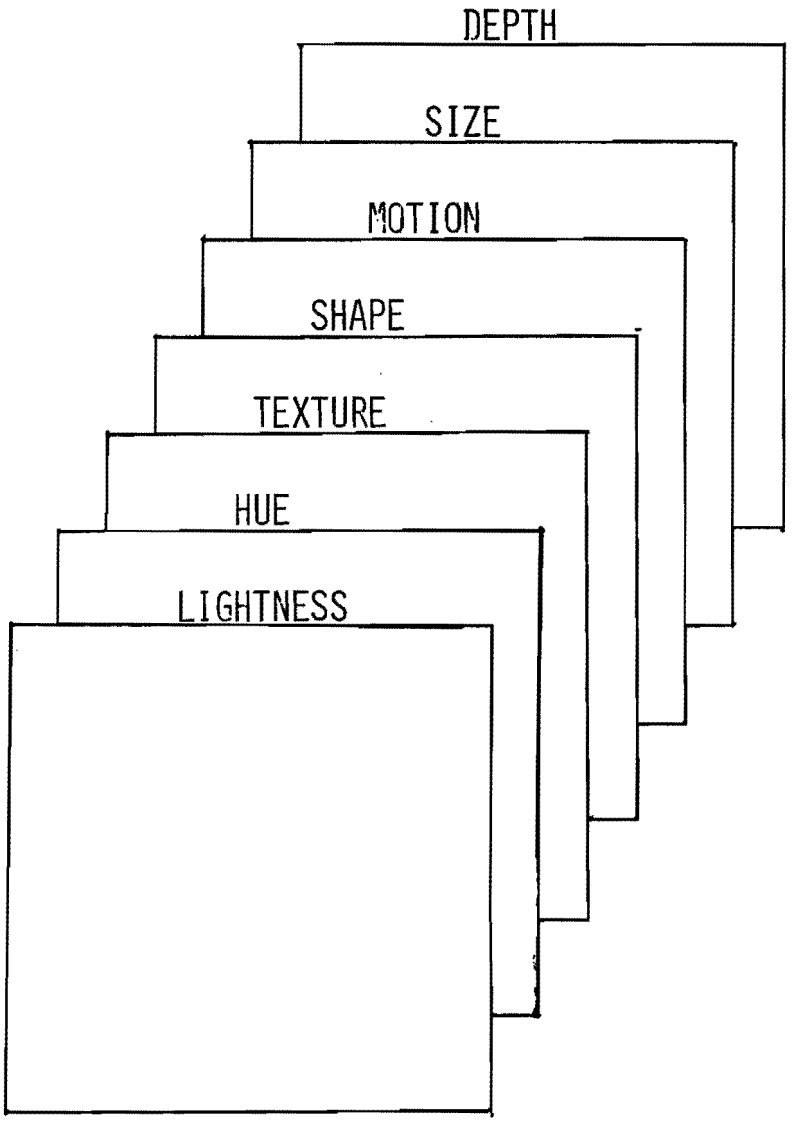


Figure 3.1: The six feature (+ depth) planes for the small world SFF

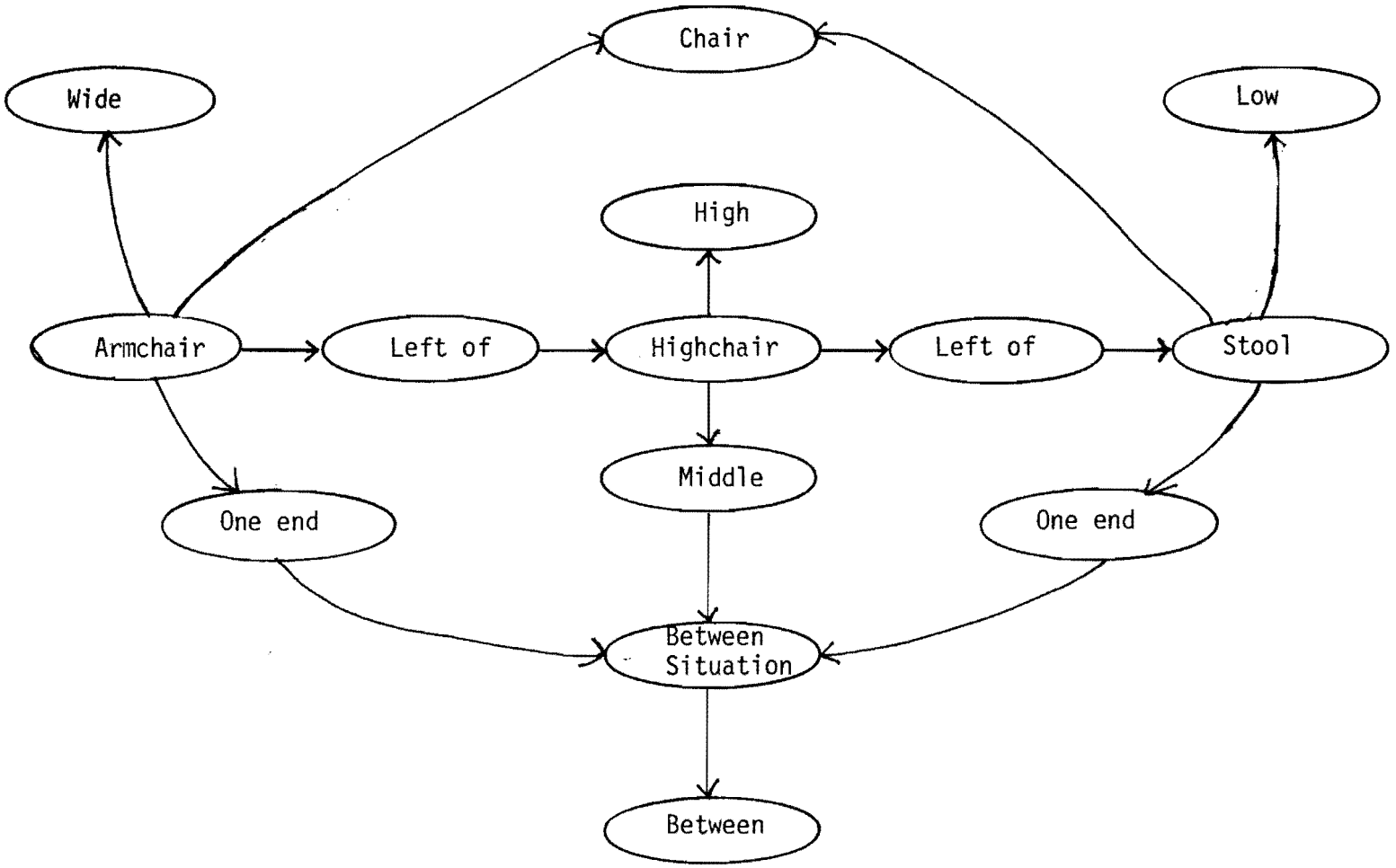


Figure 3.2: A typical network representation of visual objects in a situation
[Ballard and Brown, 1982]

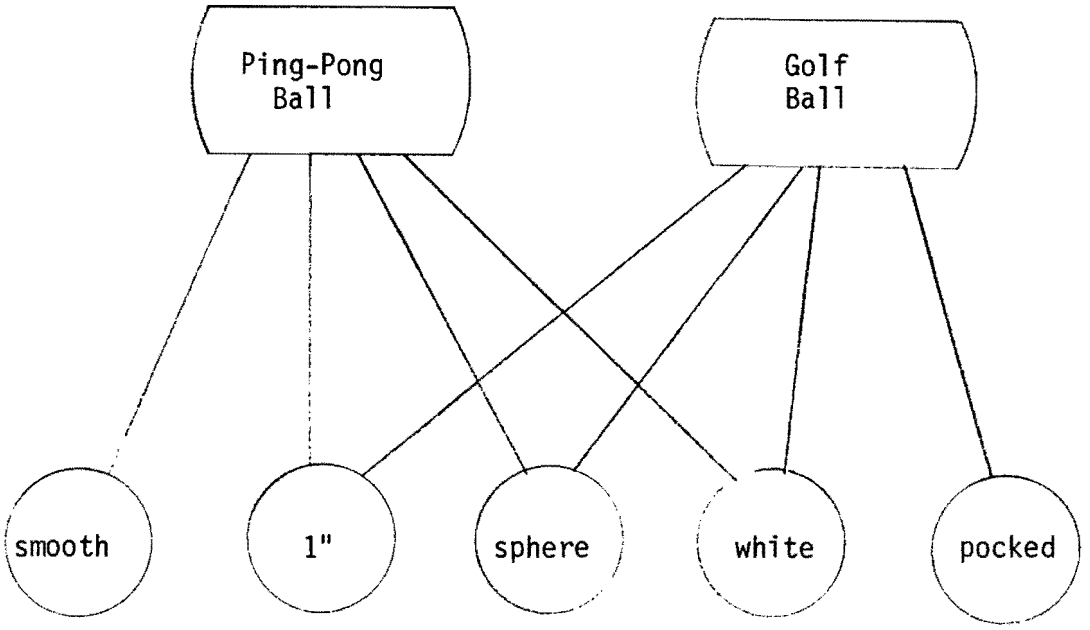


Figure 3.3: Ping-Pong and Golf Balls

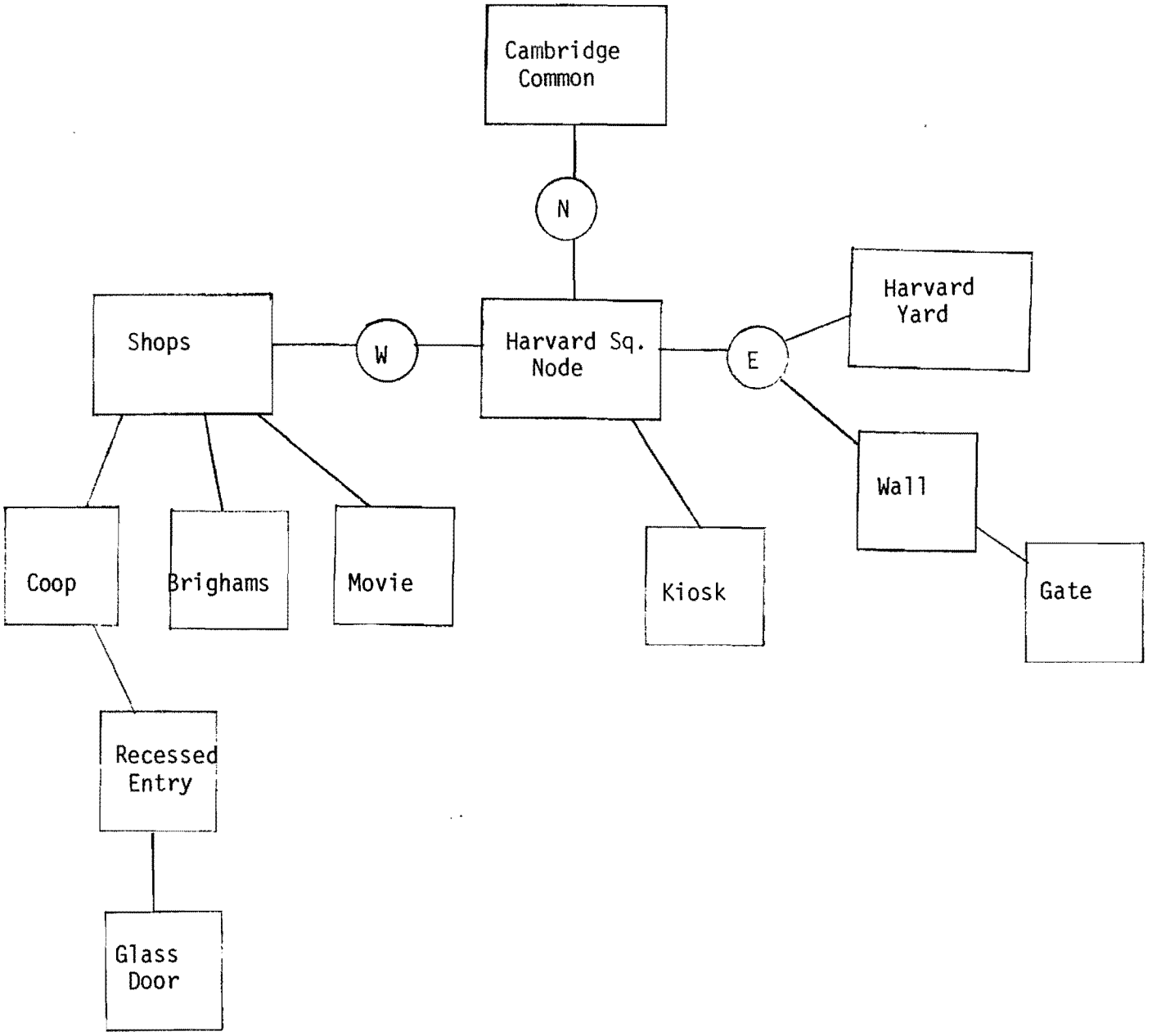


Figure 3.4: Harvard Sq. situation network rectangles are situations, squares are (complex) objects

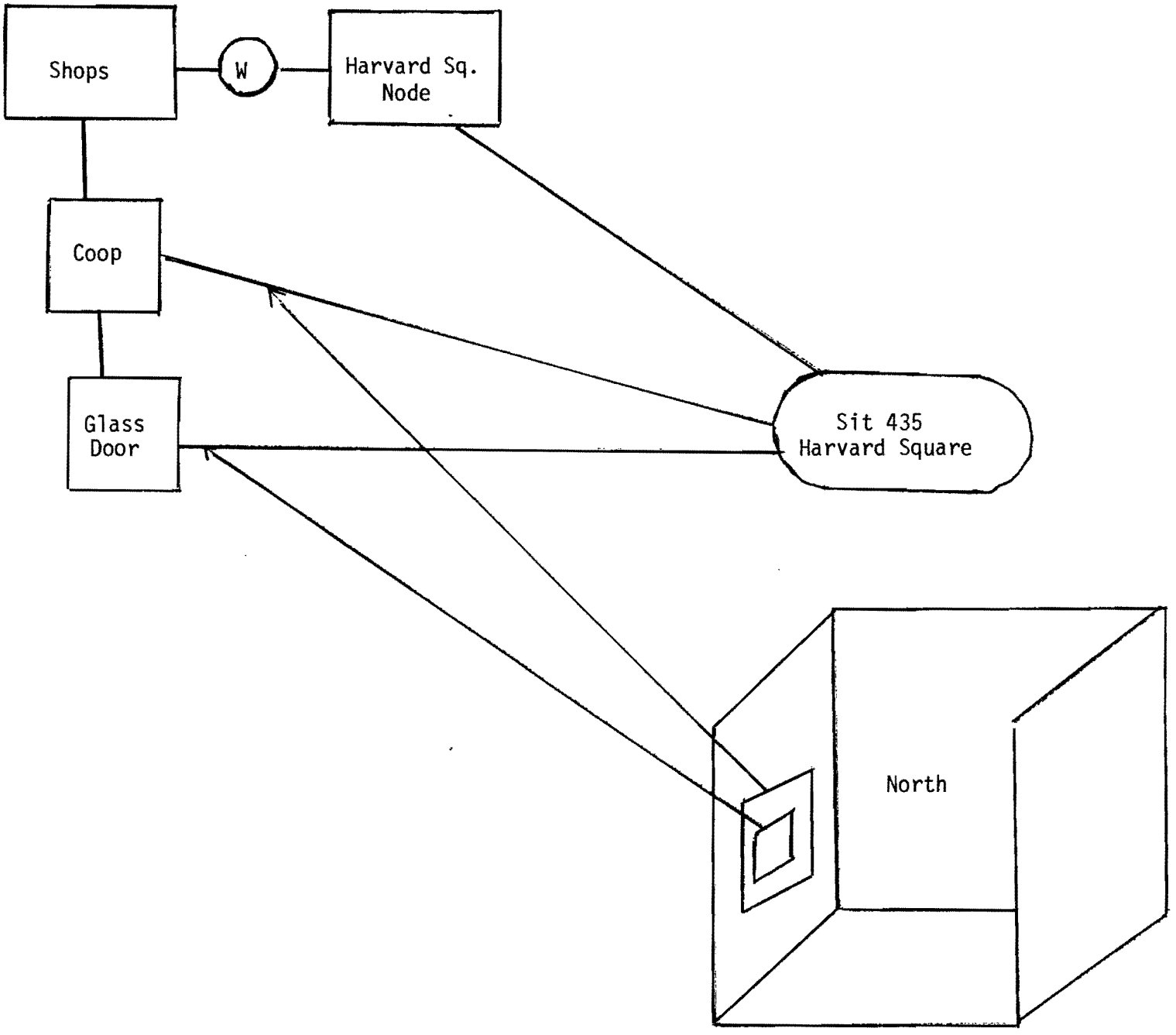


Figure 3.5: Two EF units of different scales activate different objects in Sit 435 = Harvard Square

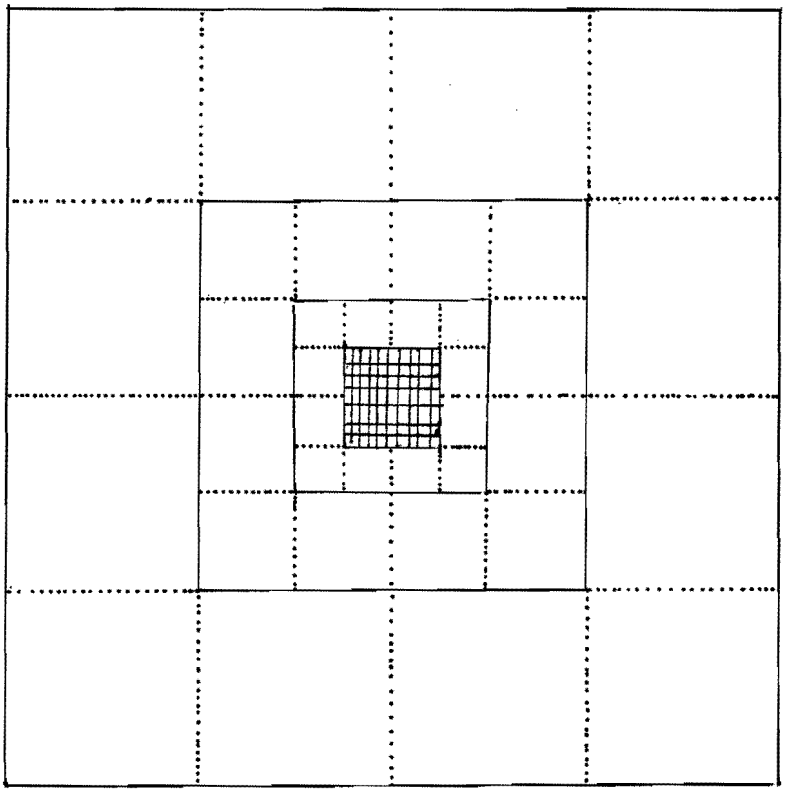


Figure 3.6: Logarithmic Retinal Frame

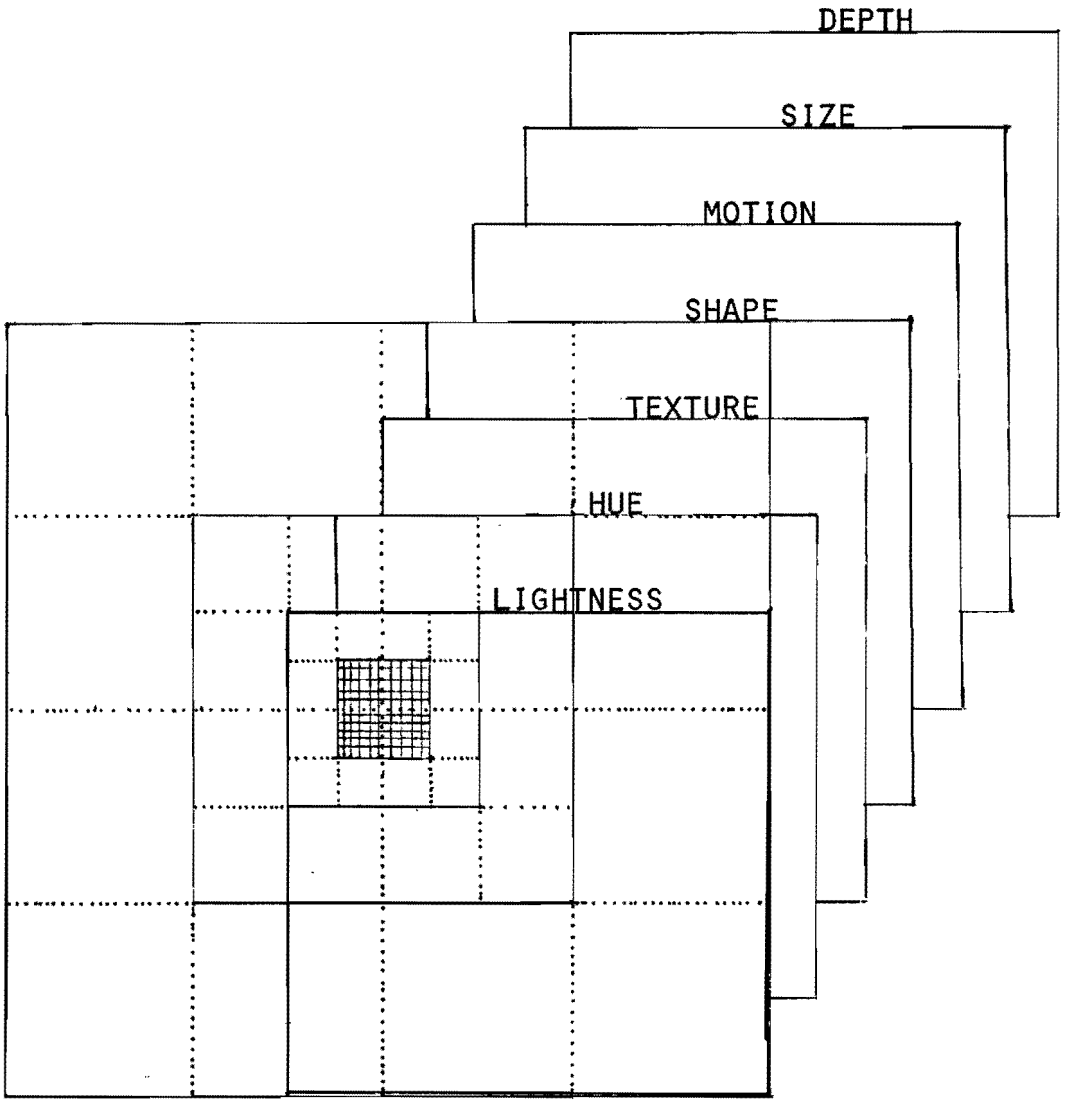


Figure 3.7: Retinal Frame mapped to SFF

4. Small World, expanded.

The purpose of this section is to address a variety of technical questions that were suppressed in the previous overview, still without seriously confronting experimental data. The technical questions are all addressed relative to the specific formalism described in Section 2, but most of the questions would arise in any attempt to model vision and space at the current level of specificity. We will follow the same order of presentation as before, but will also include discussion of some links among the four frames that were ignored earlier. Most of the specific solutions to technical problems will be carried out at the Small World scale, hopefully making it easier to see the ideas.

4.1 The SFF reconsidered.

The first technical questions concern the assumption that the Stable Feature Frame (SFF) can continuously maintain values for the hue, saturation, size, shape, color, texture, motion and distance of features in the current field of view. A large fraction of the current effort in computer vision is focussed on these problems and, while a great deal is known, quite a few problems remain. Without attempting to survey all this work, we can indicate extensions to the Small World SFF model that make it a reasonable abstraction of our current understanding of constancies (= intrinsic images = extra-striate visual maps).

There was a certain sleight-of-hand in the previous description of SFF functionality. In order to even define SFF features like shape and size, the image must already be segmented into regions, and we have not specified how this segmentation is to happen. (This is our first technical problem and is typical of the ones to follow.) Our notions of how region analysis and feature extraction are cooperatively computed is described in detail in [Ballard, 1981]. The basic idea is that the SFF also contains parameter space networks representing the relative importance of different feature values in a given scene. Color is a particularly easy example to examine. Our ten values of hue and lightness yield 100 color values that could be present in a scene. Imagine one unit for each of these 100 values whose activity is a measure of how much of this color is in the scene. Now consider the most active color and the points in the SFF whose hue and lightness yield that color. This collection of identically colored points is a good candidate for a meaningful region, especially if the points are adjacent. If there is no significant variance in depth, texture, or motion over such a region, it will almost certainly be segmented out and its size and shape can be computed. When the various features do not agree, people have trouble with segmentation (e.g. camouflage). Algorithms for forming distinct regions within a cellular computer like ours are not trivial, but are in the literature [Minsky & Papert, 1972]. The size and crude shape of an identified region could be calculated by a parameter network [Ballard, 1981]. We assume that for indexing, the properties of a region are represented by the unit at its center of mass, with the other units reporting null values.

Current Computer Vision research is directed at a slightly less abstract set of constancy features emphasizing e.g. local surface slant and tilt instead of our shape features. There is no reason why the SFF could not incorporate multiple levels of features and we expect that it will have these as well as global parameters such as the direction and color of illumination. The model also should be refined to account for the fact that there are order relationships among the features. It turns out that depth precedes lightness [Gilchrist, 1977] and that region properties like size and shape presume some segmentation by depth, color, motion and texture. All of these

calculations do interact with each other as well as the with the bidirectional (indexing and context) links to knowledge of the appearance of objects (WKF). The model presumes that this giant network relaxes into a consistent stable coalition (Section 2) and preliminary simulations [Sabbah, 1981] are encouraging, but a great deal of work remains before we can have real confidence in the computational stability of the model.

Another important issue is the role played by points of discontinuity (edges) in the SFF. Both the behavioral and physiological data indicate strongly that the visual system responds primarily to differences (e.g. in color), but the SFF encodes point values of features. The model uses the SFF primarily as a buffer memory and for indexing - both functions are better served by attempting to capture the (constancy) values of visual features. It might be useful to add additional planes representing, e.g., depth discontinuities, to the SFF and there is no problem in doing so. Depth discontinuity points would be particularly useful in grouping regions into separate objects and this, in turn, would greatly simplify indexing (which is a major technical problem to be addressed below). More generally, the conversion from retinal (difference) information to SFF (constancy value) information is a major prediction of the model. The model postulates that the SFF continuously computes, among other things, smooth continuation values for feature plane units not foveated recently.

In Section 3, we described the RF → SFF mapping as involving moving the logarithmic retinal frame over the SFF spatial map. The next task is to show how this is accomplished using the mechanisms of Section 2. The same idea of a variable mapping will occur repeatedly below. All of our variable mappings will rely on conjunctive connections; the particular scheme for the RF → SFF map is shown in Figure 4.1. First consider the case where a position in the SFF is currently covered by a equal size piece of the RF. For example if gaze were directed to its maximum extent in the upper right corner of the field (8,8), then the SFF units at position (6,5) would get values from the RF unit (64) in the spiral numbering order. This is shown in Figure 4.1 as a conjunctive connection on the (6,5) unit of links from [gaze = (8,8)] and RF position = (64). The same gaze value maps RF position (73) to SFF position (9,5), and so on. Also shown is one of the 64 other conjunctive inputs to the SFF (6,5) units; this for gaze (7,8). The mapping for unequal sizes of RF and SFF fields is only slightly more complicated. Coarse RF units map the same value to several SFF units. Fine RF units would have to compute some summary value of their findings, for each of the seven planes of the SFF. There is no difficulty here in mapping, but the nature of the RF foveal computations and their use is a technical question to which we will return in Section 4.4.

Figure 4.1: Mapping retinal to SFF coordinates, Detail

Another general issue is the choice of one unit per feature value as a basis for representing information. Although this unit/value principle is a convenient way to build models and appears to be a reasonable abstraction of the experimental data, the real situation is more complex. Even on pure computational grounds, it is much more efficient to use some encoding tricks such as the coarse-fine coding trick described in Section 2. These tricks also exploit conjunctive connections to reduce by a large fraction the number of units that would be required to capture a given level of precision for a feature value. The assertion here is that these technical tricks are

sufficient to solve the problem of combinatorial explosion in the number of units as we move to realistic numbers. Our exposition will continue to employ pure value units (e.g. in the planes of the SFF) with the understanding that any physiological predictions would have to be translated to realisitic encodings.

4.2 Indexing and Context Mappings

In this section we attempt to confront a complex set of interacting technical questions upon which the viability of the provisionary model will stand or fall. The crucial issue is how to convert from a spatial, visual, syntactic representation to the more general, modality-independent semantic network which is claimed to embody one's world knowlege. Esssentially the same problem arises in any formulation and our attempted solutions may be of some heuristic value even if the four-frames model turns out to be useless.

Recall that Section 3 presented a simple and direct model of indexing from visual features (SFF) to visual primitives (WKF). A primitive was simply any node (~ unit) in the WKF which could be indexed by a vector of feature values. Although it was not stated explicitly, the implication was that conjunctive connections would be used to activate the visual primitive when the appropriate feature values all appeared at the same point in space (and thus in the SFF). More complex objects and situations were assumed to be built up recursively from primitives using standard relationships (e.g. "below") from semantic network theory. In addition, context links from the WKF to the SFF were supposed to prime certain feature value units from general knowlege and expectations. The remainder of the section lays out how the model does all these things without attempting to specify the details of semantic network representation in the WKF, this being a major intellectual problem, independent of vision and space.

The classic problem in parallel models of indexing is cross-talk or confusion of features. If a red circle and a blue square appear together, how does the parallel network avoid activating the red square primitive? The obvious way to handle the red-circle, blue-square problem is to have a red-circle conjunctive unit for every position in the visual field. This quickly becomes infeasible for more complex combinations of features. For example, in the Small World with six 10-valued features, one would require a million units for each position in the SFF in order to implement our naive notion of mapping from visual feature vectors to visual primitives. For realistic numbers the problem grows too fast for our coding techniques [Feldman & Ballard, 1982] and other ideas must be invoked. The particular solution used here to the feature-cross-talk problem will be presented in some detail, both becuase of its importance and as an indication of how the elaboration of the model is proceeding.

The basic idea is to maintain spatial coherence for all *pairs* of property values and to index use conjunctions of pairs. Figure 4.2 depicts the basic situation for a golf ball in the Small World. We assume for now that the appearance of a golf ball is characterized by exactly one value for each of the six visual features, appearing together at a point in the visual field (SFF). There are 15 (5 + 4 + 3 + 2 + 1) ways of making pairs of values from six features, any subset of which could be used for indexing. Suppose we just use shape conjoined with each of the others, yielding five pairs involved in the indexing of golf ball appearance. The important point is that the feature-pair units are all spatially independent; there is only one white-sphere unit. The feature-pair units are themselves activated only by the simultaneous appearance of their component features at the same point in the visual field (Figure

4.2 shows size and shape at (1,6) in the SFF). For the Small World, this would mean 100 conjunctions of two inputs each to feature-pair cells. If all 15 pairs were laid out, there would be 15 x 10^2 or 1500 pair units because each element of each pair could have ten different values. Even counting the 100 separate input sites to each of these pair nodes as a unit, one gets only 150,000 units as opposed to the 100,000,000 needed for directly encoding each vector of 6 feature values at each position. Since each feature pair unit responds to the entire visual field, the model automatically generalizes from an object learned at one spatial location.

Figure 4.2: Indexing and Priming, Detail

What price do we pay for this dramatic reduction in unit count? The main cost is an increase in the chances of false indexing, the feature-cross-talk problem with which we began this section. While each feature-pair is required to be spatially coherent, the pairs could all come from different parts of space. For example, if an orange at (4,7) and a flying ping-pong ball at (1,6) occurred in the same image, the network of Figure 4.2 could falsely activate golf ball. In a more complete version with all fifteen pairs, several pairs (pocked flying, pocked white, pocked 1-inch) would not activate and this might be enough to prevent falsely activating golf ball. Other factors include mutual inhibition by ping-pong-ball and the effects of the situation context, but there remains a possibility of false activations through coincidence. In fact, just this kind of cross-talk is found in [Treisman, 1982] One cannot effectively index the entire scene and must use fixations and internal focus of attention to deal with things sequentially. Changes in region grouping and problems like transparency also require sequentiality.

There are also some minor technical questions to be answered about this scheme. One obviously must allow for indexing by more than a single value of various features. There are two cases, both of which fit quite well with other aspects of the model. When a range of values (e.g. lightness) is possible, we assume indexing is done with a coarse-valued cell which we need for other reasons anyway. If no values of some feature are criterial (e.g. hue of jelly beans), that feature is simply not used in indexing. Also, the disjunctive input sites of Section 2.2 provide a natural way of encoding separate visual appearances of a single primitive. The hard problem is how all this structure could get built for new objects, and this will be treated fairly carefully in Section 4.5.

Once an object instance has been recognized, it has a representation in the current situation independent of whether it is currently in view. For top-down context mapping to be effective, there must also be a link from visual primitives in the WKF to their component features in the SFF. Assume that the links without arrows in Figure 4.2 are bi-directional. Then anticipating the appearance of a golf ball would prime all the appropriate feature-pair units (e.g. 1" sphere). The feature pair units could, in turn, prime the appropriate feature-at-position units (e.g. sphere at (1, 6), 1" at (4, 7)). This would give some advantage in the WTA competition at each point to anticipated features but could not be very effective because it would be identical across the visual field (SFF). A much more powerful context effect can be achieved by adding spatial focus units depicted as a diamond unit in Figure 4.2. Each spatial focus unit could conjoin with context links so that only the anticipated feature-at-position units were primed. Spatial focus has been shown [Feldman & Ballard, 1982] to be a general solution to many cross-talk problems and appears to be

related to attention [Posner, 1978; Treisman *et al.*, 1980]. The coordination of spatial focus with the action of the RF will be discussed in Section 4.4.

Meanwhile, for spatial focus to be feasible, one needs a mapping from the instance (hexagonal) nodes of Figure 4.2 to the spatial focus (diamond) ones. Such a mapping encodes the (rapidly changing) information that some object instance is currently at a particular position in the visual field. This is just the kind of mapping for which the uniform connection networks of Figure 2.4 were developed. Once the links are established, the activation of either a spatial position or an object instance will strongly prime the partner. It is also not difficult to augment the spatial focus network so that the expected position of visible objects after head movements can be primed. For both computational and scientific reasons, the current model assumes that this expectation is done for only one object and the rest of the SFF is recomputed, using a little context priming but mostly direct visual input.

Complex objects (and situations) are represented in the model as networks (in the WKF) of nodes describing visual primitives or other complex objects. There are tremendous problems of several different kinds in these semantic network models and these are the subject of the next paper in the current series. Our goal here is just to provide a plausible (although crude) model of how network representation of visual appearance could fit in the four-frames paradigm.

As mentioned in Section 3, the basic idea is that each visual primitive of a complex object is represented by a node that corresponds to a particular set of feature values as computed in the SFF. Since indexing from features to primitives occurs in parallel, there will usually be several simultaneously active primitive nodes for a complex object currently in view. This simultaneous activation of subparts will tend to cause the correct complex objects to be activated, independent of the details of how the relationships among the subparts are modelled. When we consider the details of complex object representations, a number of difficult technical problems arise. This is the subject of Hrechanyk's forthcoming dissertation [Hrechanyk & Ballard, 1982], and we will be content here with a loose discussion, based on the example of representing the visual appearance of horses. Recall that the WKF visual appearance models are far from complete -- they are more like a verbal description of something not currently in view.

Obviously enough, the side and bottom views of a horse have relatively little in common. Even within the side view, the horse could appear in a variety of orientations and scale configurations and the relative positions of its subparts could also differ considerably. We must also account for the facts that there could be several distinguishable horses in a scene and that some of these may be partially occluded. Our current solution, depicted in Figure 4.3, involves instance nodes, separate sub-networks for different views and cross-referenced structural descriptions. The prototype horse has a general hierarchical description where, e.g., the trunk is composed of a body, legs and a tail. What visual primitives might be involved in recognizing a horse will depend on whether it is a front, side or other view. Thus the matching process would select together a prototype and a view which best matched the active visual primitives. Figure 4.4 shows a typical relation in the triangle notation of [Hinton, 1981]. As always, there is assumed to be mutual inhibition among competing object descriptions and view nodes. A serious weakness of the current scheme is that it has no verification apparatus for checking that the pieces of the putative horse are all connected in appropriate ways. A CM approach to the verification of the detailed geometric correspondence between a WKF model and an image is described in [Hrechanyk & Ballard, 1982]. Their solution requires an

auxiliary structure for computing the correspondence and entails a hierarchical matching strategy that is compatible with the hierarchical descriptions in the WKF.

Thus far our discussion of object recognition has been traditional in its treatment of occlusion--we ignored it entirely. We did discuss discontinuities (edges) earlier in this section and certain discontinuities (e.g. depth, motion) provide cues to possible occlusion. A more thorough treatment would include explicit occlusion-feature recognizers in the SFF, but this requires no qualitative changes. The hard problem is how to make use of occlusion cues in matching partial collections of visual features to appearance models. Our indexing scheme does not depend on totally matching features with primitives, but we need to make much stronger use of occlusion information.

The best use of occlusion information would be in connection with spatial focus and the kind of successive refinement of matching described in [Hrechanyk & Ballard, 1982]. Occlusion cues such as depth discontinuity could be used to separate areas of space believed to index separate objects and the appropriate subparts matched in the SFF. One could also add general matched-by-occlusion links to higher level nodes in the object appearance models [Sabbah, 1981]. If we are able to compute the overall position and scale (fairly accurately) of the occluded object, then the various visible pieces could be separately foveated and used to index. This is not much different than what is needed to recognize an unoccluded object that occupies a large amount of the visual field. Presumably the instance nodes recruited for the various objects could include occlusion links tied to the current situation and viewpoint. In important cases, this occlusion information could become part of the situation description.

Another major problem is multiple horses in a scene. To represent multiple horses clearly requires some kind of "instance" nodes to keep track of the positions and properties of the various horses in the scene. The model assumes that people can deal with a few instances, but must recognize (foveate) one at a time for indexing to work. Basically we assume that when a particular horse instance is foveated, the position, structure and other features are simultaneously active. The instance "node" is the set of binding units (Section 2.4) recruited to hold the coalition together. The statistics of recruiting would be between the uniform networks of Figure 2.3 and the random networks of Figure 2.4 since there is an intermediate amount of structure. The coalition representing the horse-instance-at- position could also include nodes that captured detailed orientation parameters and presumably even concepts like gait, although motion presents problems not yet solved.

The model also includes in a natural way the occurance of special nodes and structures for particular horses that one knows well. Learning the appearance of a new object, such as a centaur, involves synthesizing new structures which make use of existing substructures. Such permanent structures are presumed to arise from temporary coalitions by strengthening connections as described in Section 2.5 and [Feldman, 1981]. The model suggests that people with horse structures for particular horses, breeds, liveries etc. should be able to effectively represent more complex scenes without cross-talk. We will return to the role of network structures and foveation in the section on the retinal map (4.4). The next topic is "situations" which are WKF networks which may include several complex objects.

4.3 Situations and the EF

We are, again, tracing around the four-frames diagram of Figure 2.1. Recall that the Environmental Frame (EF) is postulated to be the multi-modal representation of the objects in the current situation. As was the case with complex object networks, the WKF network representing a situation will be more like a verbal description or sketchmap of something not currently in view. The nodes of a situation network represent either objects or sub-situations, in exact analogy to the networks for complex objects. The situation networks are assumed to be oriented by compass direction and to contain some distinguished objects that serve as landmarks. Situation networks can be conditionalized on points in time or seasons of the year.

We assume for now that only one situation is active at a given time. Since the active situation network is a stable coalition, all of the object and sub-situation nodes are also active to varying degrees, providing top-down context to perceptual processes. So far, this presents no technical difficulties; the problems arise in relating the current situation (in the WKF) to the hypothesized spatial frame in the EF.

Recall that the EF was assumed to be organized as units representing fixed positions in space. The EF is organized around cardinal directions which we call N,E,S,W and Up and Down. The model suggests that this spatial frame does not necessarily change with body movements; it is an allocentric rather than egocentric representation. The position and orientation of the ego within the EF is also maintained at all times and used in directing actions. Conceptually, one would like to be able to map the current situation network (from the WKF) to the EF such that each landmark object is mapped to its canonical position. This would enable the model to anticipate what should be seen at different positions and scale values in the environment and where to look for expected objects. Technical problems arise in trying to lay out these WKF-EF mappings in a way that has plausible resource requirements and is resistant to cross-talk.

The basic form of our technical solution is shown in Figure 3.5. The central idea is to use special situation nodes (depicted as ovals in Figure 3.5) to bind together the mapping from a fixed place unit in the EF to object units in the WKF that are expected at that place in the active situation. For reasons we will get to later, there is no link from objects in the WKF to their positions in the EF. Conjunctive connections link a position in space, represented by an EF unit with a particular object node in the WKF. When a particular situation node (e.g. Harvard Square) is activated, then activation of a particular EF node (East, Middle distance) could lead to activation of a node in the WKF representing a middle distance view of the Harvard Coop. The model assumes that the amount of EF → WKF activation is related to foveation and attention. There are also implications for retinal (RF) mappings which we will discuss in the next section.

There is a nice correspondance between the hierarchical situation representations in the WKF and the EF representations of space at different scales. The expected view of a landmark object in a situation depends on both the direction of gaze and the computed position of self relative to the EF. Moving close to an object of interest could lead in a natural way to switching activation to a sub-situation which has a more detailed view of the object. The model thus suggests that situation nodes are arranged in a discrete hierarchical structure, and that changes of visual context are discrete. In addition to scale change, other reasons for changing the (unique) currently active situation include moving out of a situation or passing a particular landmark [Kuipers, 1973]. We also assume that a change of internal focus of

attention is usually accompanied by a switch in active situations. The model can also accommodate scenarios (time sequences of situations), but we will not deal with scenarios in this paper.

There appear to be no technical difficulties in the CM representation of these ideas. Counting arguments limit the number of situation nodes to a few thousand, but this seems plausible. Some situation networks are assumed to be general (e.g. office) and used when no more specialized network is available. New situations are assumed to be handled by recruiting additional binder units linking landmark objects with their EF positions, using the techniques of Section 2.4. It is this collection of binder units that we refer to as a "situation link."

The amount of and accuracy of information captured in a situation network is quite low, but this appears to be consistent with what is known about people. One consequence of the model in its current form is that there is no link from an object-situation pair to the EF node where it is expected. One could easily add these links but this would lead to vast numbers of input links to each EF node violating a constraint. In addition, these WKF→EF connections could cause confusion between what objects were being activated in the WKF and where gaze was directed. The model currently allows one to think about one situation while visually coping with a different one, as long as the non-visual situation does not evoke (simulated) spatial reasoning or action. For the model, the position of objects in a situation is represented relationally in the WKF only and one's ability to locate objects not currently in view should be crude, unless a need for recalling the location was anticipated. This is typical of the kind of crude prediction of experimental consequences which will occupy us in Section 5.

4.4 Foveation, Pursuit Mode and the Retinal Frame

The logarithmic scaling of Figure 3.6 is about all that has been specified so far about the Retinal Frame (RF). The model assumes that the RF continuously computes proximal (non-constancy) values of visual features and transmits values to the appropriate SFF units depending on the direction of gaze (Figure 4.1). Obviously enough, the RF is intended to correspond roughly to primary visual cortex which is, by far, the best understood of the four frames. We will consider in Section 5 the evidence on what the units in primary and secondary visual cortex compute and whether RF-SFF distinction makes sense of the data.

For this section, the crucial questions are computational. One computational refinement that is required is that units in the RF can not be assumed to respond to only one feature. As we have seen, units that respond coarsely along some feature dimensions and finely along one dimension have computational advantages and we assume that this is the nature of RF units. More difficult problems arise in specifying computationally how the direct measurements of the RF can be translated to the features postulated for the SFF. Let us consider motion, which is probably the most difficult case.

For RF units in a static eye, motion is indicated by "retinal slip" - a systematic change in input among neighboring units. It is not, a priori, obvious that this local information is enough to determine the object motions and light changes that could cause the retinal changes. Recent research in our lab and elsewhere [Brady, 1982] has shown that these "optical flow" calculations are feasible under a range of conditions sufficiently general for the purposes of the SFF model, which is not hypothesized to be perfect. The other SFF features -- hue, lightness, size, shape and surface texture

are assumed to be computed cooperatively from RF measures of local detectors of orientation, motion, spatial extent and disparity with different spectral tuning. The details of how the RF-SFF computations are specified is a major part of current research in computer vision [Ballard & Brown, 1982]. The totality of this work is sufficiently advanced to give us confidence that these computational issues will not be a major hurdle. Whether or not any such algorithms are used by nature is a primary experimental question raised by the four-frames model and Section 5 will be largely concerned with this issue.

There are some other purely computational issues relating to the RF - particularly stereopsis and pursuit mode. Very little has been said so far about binocular vision, because the current model assigns it no great role. The SFF is assumed in the model to be cyclopean and to incoporate two RF readings and disparity information when available. The visual field covered by the SFF is partly monocular in any event. We have discussed gaze and saccadic eye movements briefly in a couple of places. The model says nothing explicit about the choice of fixation patterns although the WKF networks for complex objects and situations would presumably help direct saccades. The question we now address is how foveation effects indexing.

The basic four-frames paradigm assumes that indexing (and its inverse, context) occurs continuously everywhere in the SFF. It also assumes that indexing is "stronger" at the place currently being fixated. In Section 3, we saw that this strengthening was a combination of selective top-down activation (through the EF and situation links) and selective bottom-up activation of the places in the SFF currently mapped to the fovea. The third strengthening effect described there was the ability to use directly the more accurate calculations of color, texture, etc. achievable by the fovea. This amounts to postulating a direct RF-WKF indexing link not shown in the four-frames diagrams. Such a link would be much simpler than the one described in Section 4.2 because it would not need spatial coherence and presumably would not have a top-down context inverse.

A direct RF-WKF indexing link is also useful when we consider the "pursuit mode" of the visual system. As we saw in the introduction, it is totally different to track your finger across text than it is to read following your finger. The literature refers to the former as the pursuit *system* but we prefer the term *mode* because much of the same structure is used in both modes. Our assumption is that the system operates in pursuit mode both in tracking a moving target and while the observer is moving under visual guidance.

Obviously enough, the purpose of pursuit mode is to keep a visual target foveated despite target and/or observer motion. Pursuit is qualitatively different in the four-frames model because the accumulation of stable constancy data by the SFF can not be the same in pursuit mode as it is in scanning a static scene. In scanning, the periphery of the RF receives input from a fixed scene (at varying resolution). During pursuit, the periphery sees a rapidly changing scene. In fact there are secial mechanisms to prevent optokinetic effects in the periphery from disrupting pursuit [Carpenter, 1977]. The model suggests that certain RFF functions such as depth and 3-D motion of the target must be computed in scanning mode before pursuit. During pursuit, we assume that the primary indexing occurs between the RF and WKF refining the parameter values originally computed by the RFF. Meanwhile two other computations are active. Optical flow calculations are assumed to be continuously operating in the RF, allowing the detection of potential collisions. The WKF is assumed to continue to register (low resolution) peripheral input from the RF as best it can. The question of how much recognition (indexing) of peripheral objects occurs

is assumed to be one of attention; if the tracking task is not too demanding, some SFF⇄WKF computations can be fit in. Such computations interfere with the convergence of the tracking function and are suppressed under heavy load.

When the observer is moving, the situation networks must also be brought into play. We postulate that the observer navigates by successively fixating and tracking landmarks. Again, peripheral vision and the SFF can do some recognition if the tracking is not too demanding. Peripheral vision, prior knowledge and occasional scanning-mode saccades enable the observer to maintain a situation network adequate to provide successive landmarks.

4.5 Learning in the CM Four-Frames Model

Acquisition of new knowledge has been the most difficult problem in the development of CM and related paradigms. Our CM model includes an assumption that there is not enough growth of new connections to account for adult learning, and changing of weights must suffice. The problem becomes particularly acute in the current context, because we must model the continuous play of transient information on the WKF as well as the incorporation of some of the information into permanent structures. The basic idea is to exploit the fact that randomly connected networks can essentially always be made to capture the required information using only weight-changing.

The current model assumes that the basic structure of the Retinal (RF), Feature (SFF) and Environmental (EF) frames are genetically and developmentally determined and do not change in normal learning. In particular, the coherence of the spatial representations and the mappings between them are assumed to be in place. In this case, most learning takes place in the World Knowledge Formulary (WKF) which encodes the observer's knowledge of the particular objects and situations that it has encountered. One must also learn the indexing - context links between the SFF and WKF and have a way of recruiting situation links to relate the EF to situations in the WKF. A more realistic model would include some plasticity in all of the frames, but the same basic considerations seem to apply.

All of the learning in the model is assumed to be accomplished by the same (somewhat magical) algorithm described briefly in Section 2.4 and more carefully in [Feldman, 1981]. The algorithm exploits the fact that large random networks have a radically skewed distribution of connections to a small subset of nodes. For example, in a graph of 1,000,000 nodes with 3000 random connections each, there will be about 29 *binder nodes* with three or more links into a set R of 20 randomly chosen nodes. If these binder nodes could be recruited properly, the binder nodes plus the previously unassociated recruiting base R would form a stable coalition. This stable coalition would be a form of coherent active memory and could serve as the basis for permanent learning of the coalition as a "concept." Section 5 of [Feldman, 1981] is concerned with describing plausible CM algorithms for all this and we assume here that the arguments there are sound.

The idea, then, is to assume that there are pools of randomly connected units available to be recruited for binders. Consider the hexagonal node in Figure 4.2. One clearly needs such instance nodes to be able to distinguish the various golf balls that might occur in a given situation. In our model, such instance nodes are recruited as being the small set of units that bind together the crucial information--here the facts that the object is a golf ball belonging to Fred in situation 67. If there were some other noteworthy fact (e.g., it was pink) the recruiting algorithm would include the

appropriate units. Usually the recruiting of a node for a visual object instance will include spatial relation links to other objects (particularly landmarks) in the current situation. We can now see that a "node" in the WKF usually consists of some binder units with connections to the various concepts semantically linked to the new "node". Instance nodes are often transient, but sometimes get incorporated into a new or modified situation description. It will come as no surprise that the "situation links" hypothesized to link positions in the EF with objects in the WKF are also randomly recruited sets of binder nodes. If a situation is deemed to be important (or importantly changed), recruiting is initiated, linking the activated objects and positions in a coalition held together by the binding situation links. Obviously enough, a great deal more work is required on the details of these algorithms, but the general idea seems no flakier than several other aspects of the model.

Even assuming that random recruiting will do all we ask of it, there remain questions of how the detailed WKF structures get built. The central question here is the extent to which we should postulate pre-wired structures and how much can be attributed to recruiting. This is, of course, the nature-nurture issue appearing in its CM manifestation and is not something to be treated in passing. A feeling for the problem can be derived from Figure 4.3, some WKF structure for horses. It seems reasonable to me to suppose that some crude structure representing the general nature of animals (other moving things in the world) may have evolved from what the Frog's eye appears to tell its brain. The only alternative (within CM) is to assume that all such structures are learned and generalized from experience. The next paper in this series will attempt to deal more carefully with the relationship between WKF neural nets and semantic networks.

Figure 4.3: General views of horse

Assuming that the SFF structure and the basic structures of objects in the WKF are understood, the index-context mappings fall out nicely. Consider the detailed golf-ball mappings in Figure 4.2. The built-in structures are assumed to include all the round and diamond-shaped nodes and their connections. The general golf-ball node is seen to be recruited as a binder linking the appropriate property-pair units with units reprsenting other aspects of golf-balls and their place in the universe. The random recruiting process specifies that the binder links be bi-directional, so that indexing and context should work as suggested. Extending all this to complex objects like the horse of Figure 4.3 appears to be feasible, especially if we assume some pre-wired structure. The point of all this is to provide a crude base for the claim that the four-frames model is not obviously wrong. The final section examines the claim a little more carefully in the light of a variety of experimental findings.

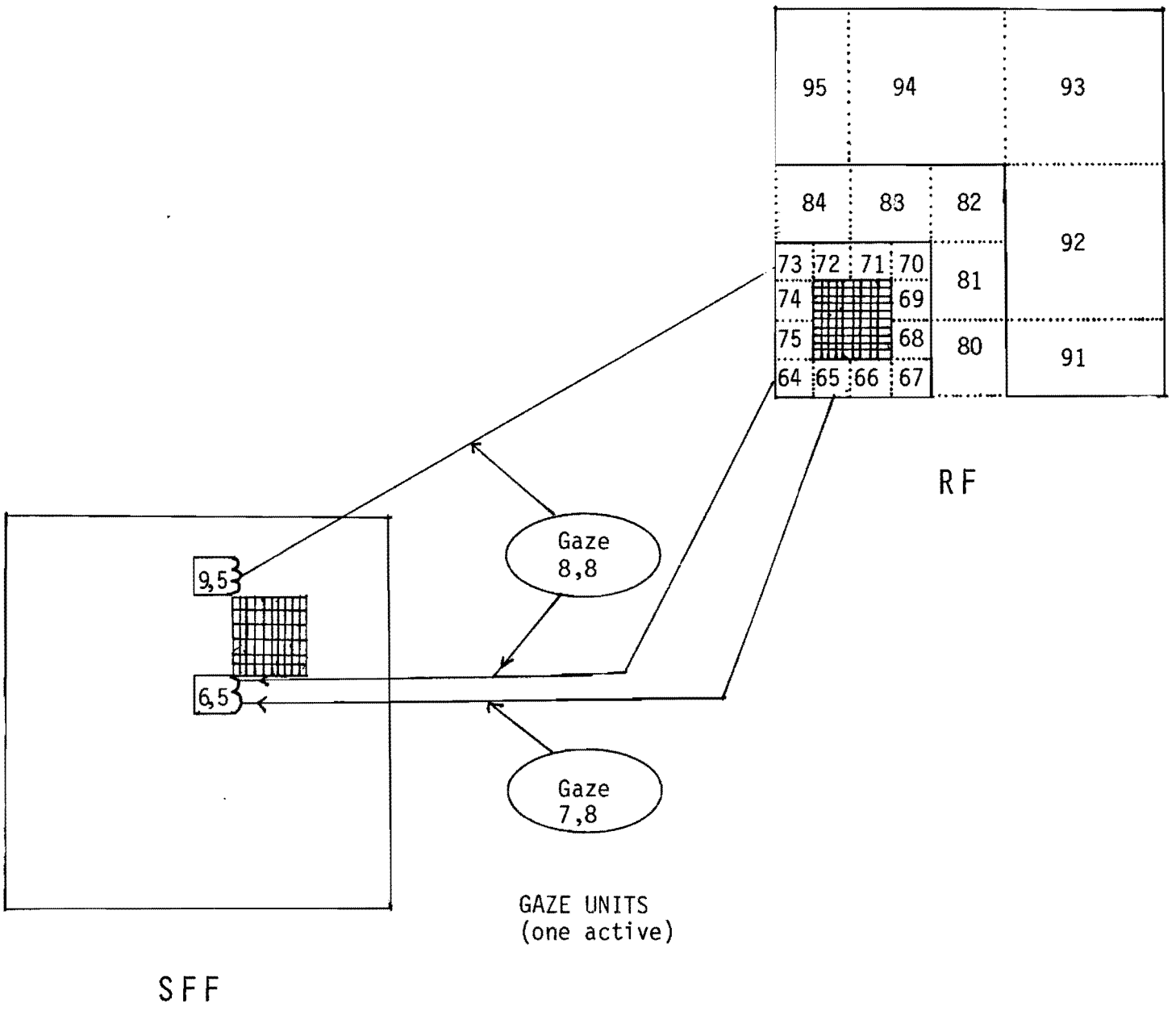


Figure 4.1: Mapping Retinal to SFF Coordinates, Detail

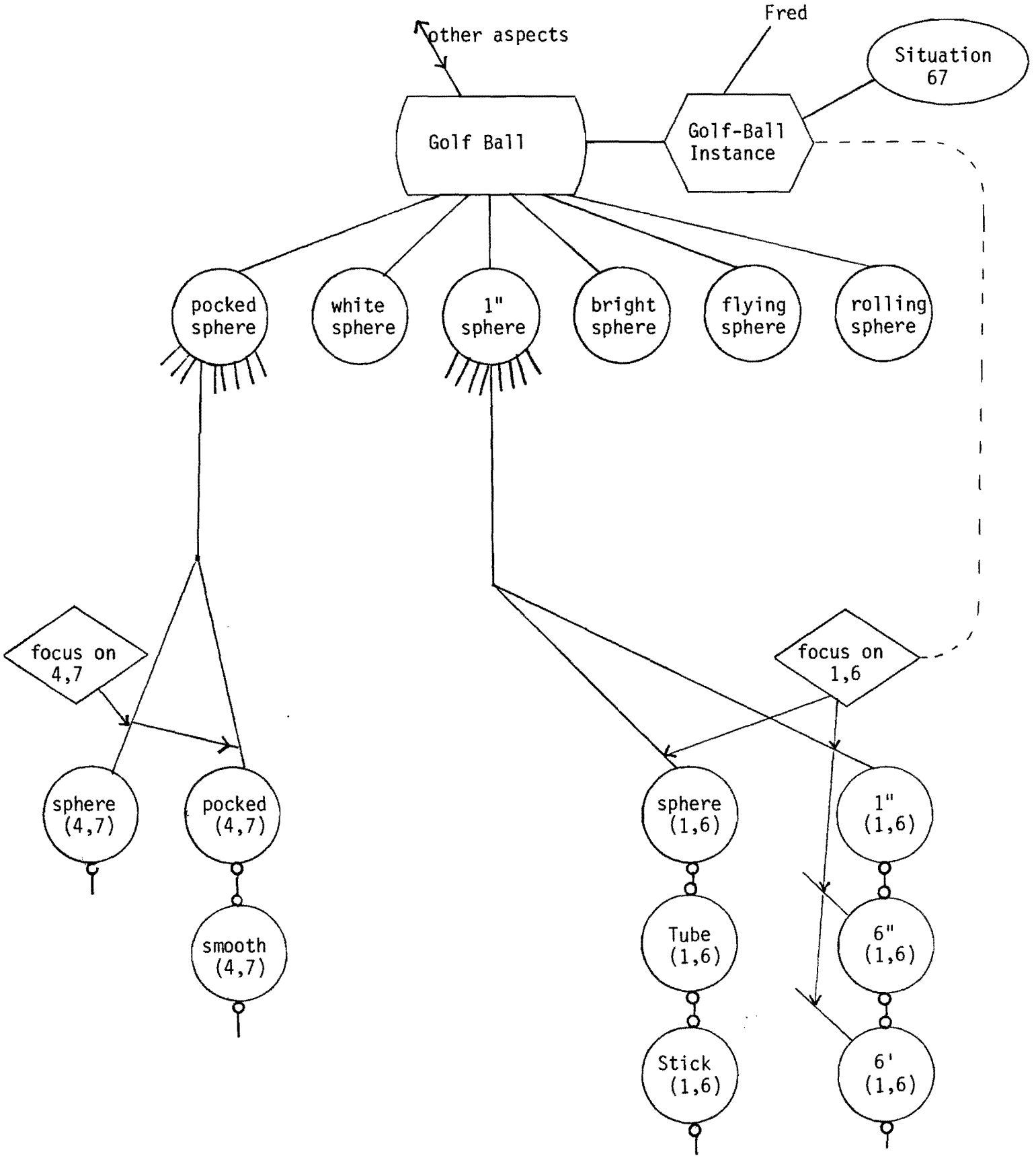


Figure 4.2: Indexing and Priming, Detail

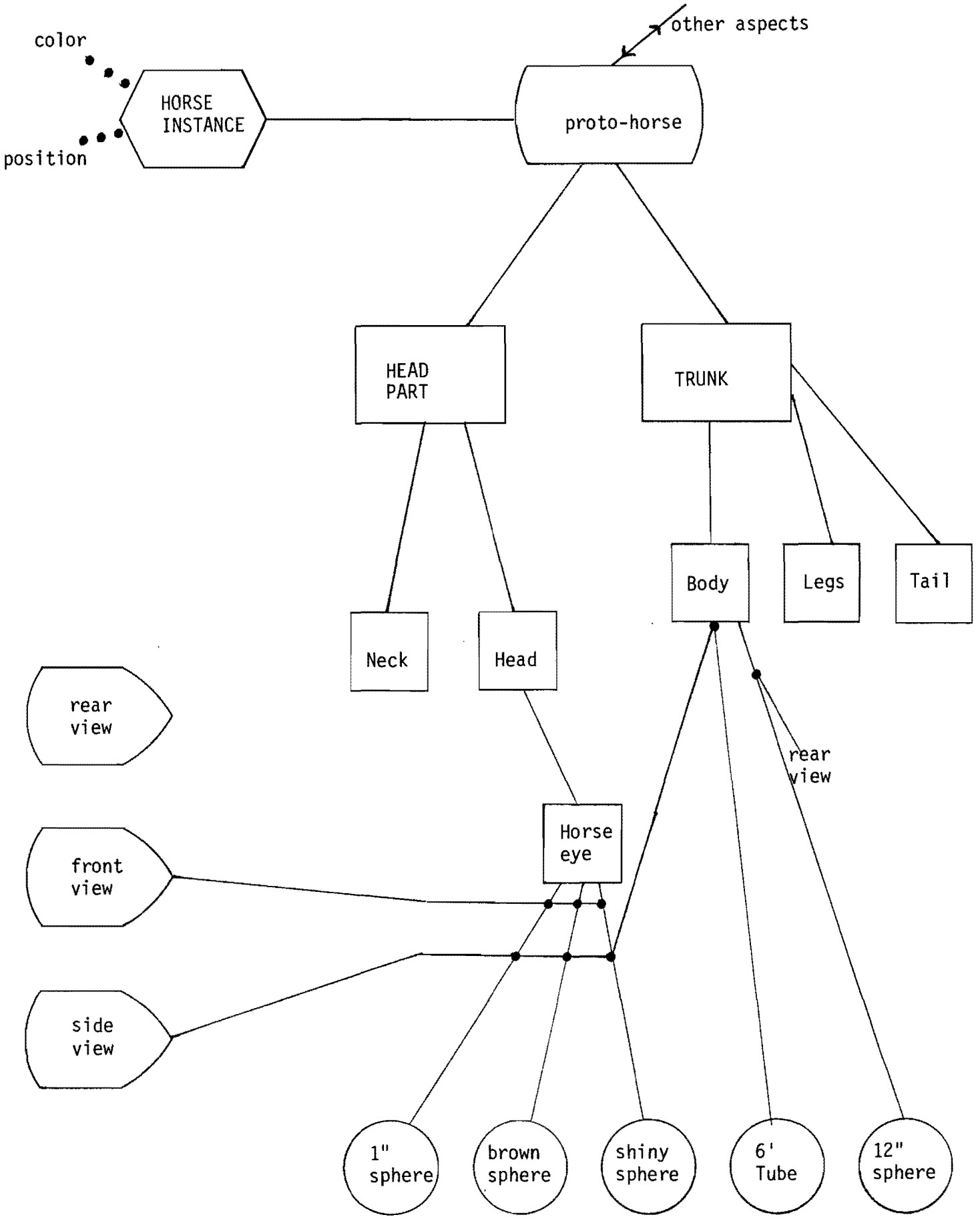


Figure 4.3: General views of horse

5. The Small World and the Real World

The major claim made for the Four-Frames Model is that it is consistent with all the established facts about vision and space. It will now be clear to the reader that the claim is, at best, a qualitative one; no particular systems or range of phenomena have been modelled at a scientifically adequate level of precision. The purpose of this section is to explore the qualitative adequacy of the Four-Frames Model and to describe some of the experimental results that led to its current form. Not surprisingly, I am currently unable to perceive any experimental results that do not fit within the model and need to have them brought to my attention.

One of the basic criteria used in the formulation of the model is that it be intuitively plausible. The discursive presentation of the four frames in the introduction is also intended to suggest why the choices are reasonable. We make no further appeal to intuition here, but would be interested in reports of intuitive dissatisfaction with the model.

The current paper arose out of an attempt to specify more precisely some aspects of the connectionist model of visual memory described in [Feldman, 1981]. We first had to develop a technical language for specifying connectionist models and learn how to use the language on non-trivial problems [Feldman & Ballard, 1982; Sabbah, 1981]. Before taking the formalism too seriously, I also had to convince myself that it was capable of incorporating short- and long-term change [Feldman, 1982]. This formalism, outlined in Section 2, has been stable for some time and is also being used in a variety of other tasks [Small, 1982; Hrechanyk & Ballard, 1982]. Its role here is to support detailed computational/anatomical representations of the various processing functions hypothesized for the model.

The behavioral and neurobiological constraints on the model were chosen as broadly as possible. I deliberately attempted to incorporate only the least controversial and best established findings. This decision fits well with the relatively abstract level of the current model. It should not require delicate experiments or arguments to point out structural flaws in the Four-Frames model. Some potentially revealing experiments will be suggested later in this section. It is, of course, enormously easier to suggest experiments than to carry them out. The main purpose of this, or any other model, is to help suggest questions that are worth the experimental effort.

Many of the elements of the four-frames model will be easily recognizable to workers in AI. The Stable Feature Frame has much in common with Ballard's parameter networks [Ballard, 1981] which is itself an extension of the intrinsic image notion which is currently a major topic in Computer Vision. The active semantic net of the World Knowledge Formulary fits into almost any current knowledge representation scheme in AI or cognitive psychology. The Environment Frame and situation links are also quite like the AI models of space [Kuipers, 1973; McDermott, 1980] to the extent that they have been worked out. The reason for mentioning all this here is to suggest that the basic computational paradigms selected for the four frames are consistent with current mainstream AI notions of how these functions can be accomplished. The translation to CM terms is only partially specified in this paper, but there should be enough material to indicate that the standard AI structures and algorithms are expressible in terms of neuron-like computing units in a way that is compact and fast enough to be plausible.

There are two lines of computational experiments that might be added to the work already underway. The small world system could be simulated as specified. The performance range would be limited but one could learn quite a lot, especially from the SFF-WKF interactions. One of the nice features of the model is that it solves the

old AI problem of converting from numerical to symbolic representations of a scene. A second line of experimental AI work could focus on situation maps and the EF. It would be very informative to see if hierarchical and sequential situations could be implemented and whether multiple situations could be worked out computationally.

But it is not computational experiment that is most needed at this stage. The Four-Frames model makes a number of predictions which should be behaviorally and physiologically testable. Computational requirements have played an important role in the development of the model, but major constraints have come from the structure and behavior of the visual system. Most of the assumptions in the four-frames model are part of a widely shared current world view and are not being explicitly addressed. What does need more discussion is the rationale for the choices made in the novel integrative aspects of the work. The experimental basis for our choices is in no instance compelling; more research needs to be done in all of these areas. Various experimental findings suggesting the central features of the four-frames model are presented as suggestive.

For the retinal frame, the data is greatly ahead of the model and the theory has relatively little to offer experimentalists. There are some new questions to be asked, but they are mainly concerned with the relation between the RF and the SFF. The four-frames model assumes that the detailed calculations of color, texture, and so on, are carried out by the RF and integrated by the SFF. We assume that striate cortex and the various psychophysical "channels" are at the RF level. Obviously any foveal functions are part of the RF. Most of [Marr, 1982] is concerned with RF calculations; he suggests a number of experiments that would also be of interest here. The most interesting prediction of the model concern the interactions between the RF and the (hypothesized) SFF. One would expect mappings to extrastriate cortex that depended on gaze, and mapped RF units with similar response characteristics. Figure 5.2 suggests that at least the gaze information of Figure 4.1 is available for this mapping through the LP-Pulvinar complex (cf. also [Graybiel & Berson, 1981]).

The Stable Feature Frame is a major prediction of the four-frames model. It presents a computationally plausible and relatively well-specified theory of the functioning of extrastriate visual cortex. It is well established that there are reciprocal connections among most extrastriate visual areas (Figure 5.1) and that the features to which each area is most responsive vary [Allman *et al.*, 1981; Cowey, 1982]. There is some evidence that extrastriate visual maps are concerned with constancy features [Zeki, 1980]. Experiments like those of [Mays and Sparks, 1980] demonstrate that saccades are directed towards points in space, not coded as relative displacements from the current fixation.

With one major proviso, the SFF makes predictions that are subject to immediate experimental exploration. The proviso is (as mentioned earlier) that SFF units are assumed for simplicity to respond only to a single feature. This is neither biologically plausible nor computationally efficient, which is a pity because it would make the experiments much easier.

Given that we are dealing with multi-feature units, the SFF makes strong and perhaps surprising predictions. One should find visual maps that are both spatially organized by head position (in an upright stationary animal) and that respond to constancy values of visual stimuli. These should interact bi-directionally with parameter maps that are organized along non-spatial axes; this latter hypothesis is currently being tested [Ballard & Coleman, 1982].

The obvious alternative to the SFF hypothesis is one that suggests that constancy and indexing computations are done separately at each fixation, with integration of the scene occurring only at our WKF level. The crucial question is the existence of

spatial maps that are independent of eye position. There are isolated reports of units whose properties are independent of eye movement [Schlag *et al.*, 1980; Tomko *et al.*, 1981], but the usual description of extrastriate maps is in retinal terms. However, the vast majority of neurophysiological experiments have been done on anaesthetized or fixated animals and would not distinguish retinal from spatial organization. It has also been noted that the receptive field size is much larger (up to the entire field) as one moves towards more anterior visual areas [Gross *et al.*, 1981]. Since most fixations are with 15°, the effective size of the SFF could be of the order of the receptive field sizes found in the extrastriate areas shown in Figure 5.2. Visually responsive areas more anterior than these will be discussed in connection with indexing and the WKF.

The psychological literature already contains extensive data on non-retinal (spatial) encoding of visual data and on constancy calculations [Fisher *et al.*, 1981; Epstein, 1977; Howard, 1982]. The notion that these are carried out (along with perceptual filling) by a single structure seems to be consistent with these literatures, and is certainly testable. Behavioral experiments like the masking work of [Davidson *et al.*, 1973] give some idea of the interactions of the retinal and spatial frame. In these letter naming experiments, masks were perceived to overlie the target letter that was in the appropriate SFF position, but it was the RF position that could not be identified. The experiments of [Jonides, 1982] suggest that random patterns can be integrated surprisingly well across fixations.

There is also evidence of important interactions among SFF computations. For example, apparent motion will not occur for objects which appear to be at great depth no matter what choices of retinal spacing and inter-stimulus interval [Haber, 1982] are used. There is wide range of experiments [Johansson, 1977] on the interactions of perceived depth, shape and motion, which are directly relevant. Another example is the work of [Gilchrist, 1977] showing that lightness constancy is applied only to adjacent areas of the same apparent depth. If the different intrinsic image calculations interact in the way we suggest, one should be able to predict the perceptual effects of anamolous combinations. An effort to deal comprehensively with existing illusion data would be a strong test for the model. One would also expect that higher-order masking and adaptation experiments [Weisstein, 1978] might reveal some of the encodings used in the SFF.

The main use of the SFF in the model was in indexing from its visual features to visual primitives in the WKF. The particular networks used (Figure 4.2) call for spatially independent units that respond to pairs of visual features. The most likely anatomical site for such units would be the infero-temporal (IT) cortex [Gross *et al.*, 1981]. Gross *et al.* report that units in this area are spatially independent and respond to complex stimuli and multiple features. The connections known for IT are also consistent with the model. There are apparently two processing stages between primary visual cortex (VI) and IT. The outputs from IT include ones that could embody our spatial focus units and indexing links to the WKF, which we presume to be subsumed by anterior temporal and parietal structures. Needless to say, there are alternative treatments of the relatively small amount of information known about this large area of cortex.

Indexing by spatially independent feature-pair units is only one of a number of possibilities. Treisman [Treisman, 1982] has a collection of experiments that limit the possible performance of such a mechanism in humans. She shows that, under overload conditions, subjects cannot detect in parallel targets requiring feature pairs (red square) but can do quite well at single-feature detections. Treisman hypothesizes that all feature-pair detections require an internal focus of attention (like our spatial focus), but this seems to me to be much too slow for coping with natural scenes. This

is another area in which the model is close enough to existing experiments for useful interactions.

The WKF, our network of world knowledge, is the least susceptible to direct biological experiments of the four frames. In the model, the WKF is recruited from all modalities and output areas. Its functions would be subsumed by a number of areas, presumably in the anterior portions of temporal and parietal cortex. Bulk metabolic experiments give some corroboration of this view, but all this is not much more than restating the classical notion of association areas. There is some eividence for multi-modal-feature cells of the sort required for the WKF being found in the Superior Temporal Polysensory area of [Bruce *et al.*, 1981]. Direct neurophysiological investigation of the WKF does not appear to be a promising route.

Behavioral testing of the WKF does seem to be feasible at present. There is considerable work in experimental psychology on spreading activation in semantic networks [Anderson, 1976; Collins, 1975; Smith *et al.*, 1974] and a fair amount on the perception of scenes [Hintzman *et al.*, 1981; Palmer, 1981]. The four-frames model suggests a number of experiments on priming, confounding, and other issues based on the proposed network structure of appearance models.

The cortical structure most likely to subsume the functions of the Environmental Frame (EF) appears to be the posterior parietal region [Lynch, 1980; Robinson *et al.*, 1978]. The four-frames model suggests that it is multi-modal, allocentrically organized and contains sub-structures that encode the current ego position. The EF should play a crucial role in hand-eye and other visually guided tasks. Most of these characteristics have been attributed to the posterior parietal area, but there is still quite a lot of disagreement on specifics [Lynch, 1980]. The EF is assumed to act through situation links connecting to WKF networks. There is considerable behavioral evidence that people employ relational, network-like descriptions of spatial situations [Hintzman *et al.*, 1981]. The four-frames models entails a number of specific predictions about these networks and about cortical connections between EF, WKF and gaze structures. The constraint of one-way EF-WKF is a computational one -- it seems unreasonable to have every object link to its places in the EF. The model assumes that objects in a situation are located relationally (in the WKF) rather than in absolute space [Hintzman *et al.*, 1981]. Results from child development studies could also be helpful here; it is already known that the ability to use allocentric frames of reference develops rather late [Piaget & Inhelder, 1967].

One way in which the four-frames model vastly oversimplies the visual system is in ignoring hemispheric laterality. Each hemisphere performs visual computations for the contralateral hemi-field with very little communication before the infero-temporal areas. The only systematic mapping across the hemispheres for earlier areas is of the vertical meridian, which is the border between the two hemi-fields. In terms of the model, this means that the RF and SFF are duplicated and that our spatially-independent-feature units (cf. Figure 4.2) are probably also separate but communicate across hemispheres. The WKF obviously would cover multiple modalities and hemi-fields and would represent the first fully centralized level. There are a number of aspects of external space known to be coded separately in the two parietal lobes, but we postulate that the EF is subsumed by the right posterior parietal region. The major problem for the model is explaining how early vision (our SFF) copes with the switching of inputs between hemispheres with gaze shifts. This appears to be a difficult and important issue in any account of vision and space.

Even without new experiments, there is a great deal that might be learned from trying to fit the four-frames model to existing bodies of data. Doing this at a crude level has forged the current form of the model. Subsequent efforts are of two

different kinds: detailed fitting of small segments of data and further refinement of the global model. Detailed studies are underway at Rochester on the occulomotor system, on parameter networks in extrastriate cortex and on computational models of specific SFF and WKF computations. These studies plus responses to the current article will hopefully lead to an improved and elaborated second version of the four-frames model. At the least, we would hope to direct some more attention to the global properties of the visual system, which is often treated as a large number of totally separate problems. The rationale of the whole enterprise is that it is not too early to benefit from more general considerations of the problems of vision and space.

Acknowledgements

A number of people have made valuable comments on earlier written and oral presentations of the model. Particularly useful were the suggestions of Dana Ballard, Paul Coleman, Francis Crick, Lydia Hrechanyk, Walter Makous, and David Zipser.

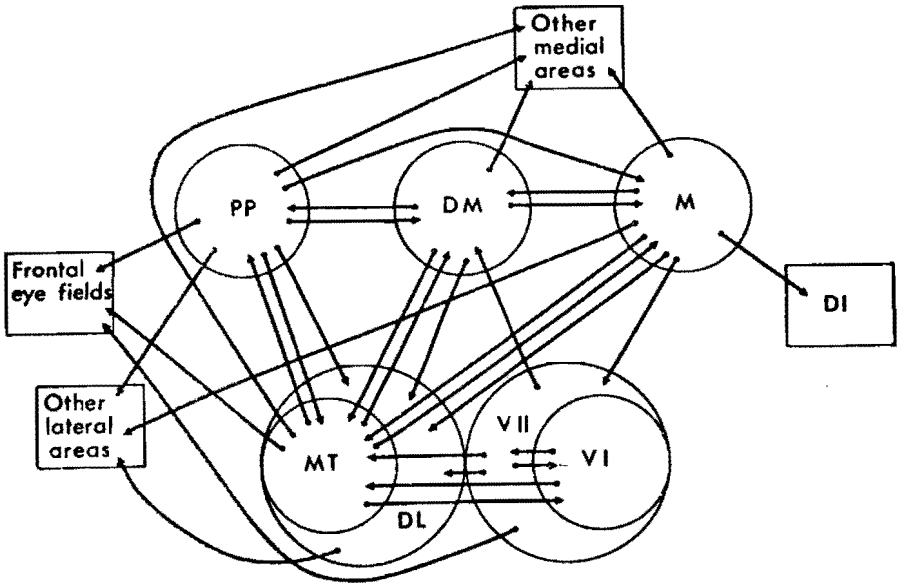


Figure 5.1: Connections among visual areas in owl monkeys. The areas are as in Figure 5.2, viz: PP (posterior parietal cortex), DM (dorsomedial temporal area), M (medial visual area, not in Fig. 5.2), DI (dorso-intermediate visual area), MT (middle temporal visual area) and DL (dorsolateral visual area). The primary visual areas are denoted VI and VII.

From: R. E. Weller and J. H. Kaas, "Connections of Visual Cortex in Primates," in C. N. Woolsey, Multiple Visual Areas, p. 137.

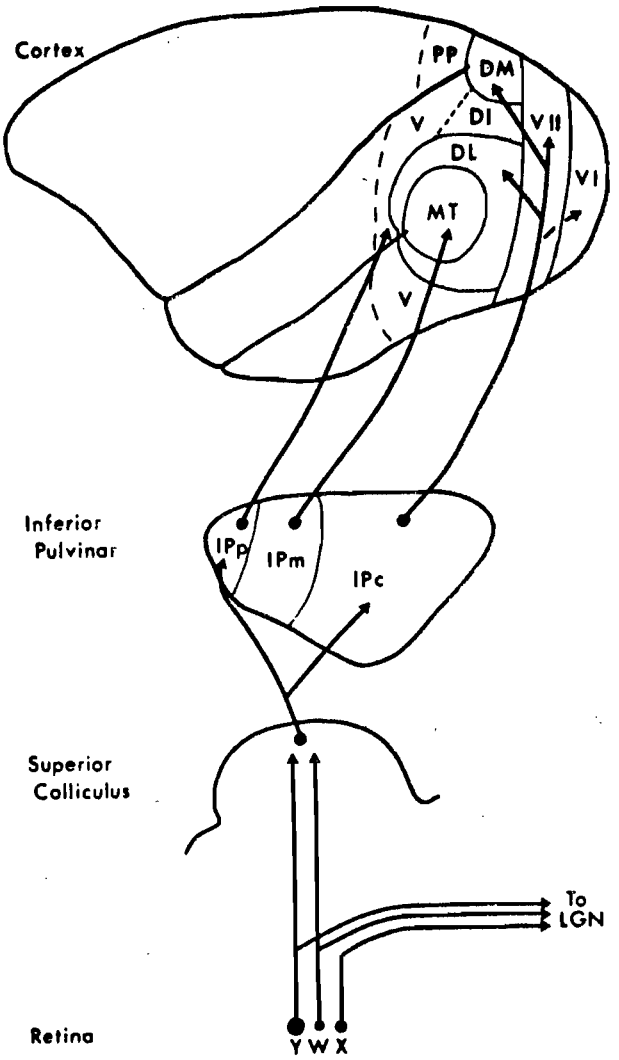


Figure 5.2: The tectopulvinar relay system. Retinal input to the superior colliculus from Y and W cells is known from electrophysiological studies in macaque monkeys. Studies in owl monkeys indicate that the superior colliculus projects to two of the three subdivisions of the inferior pulvinar complex, and that each subdivision of the inferior pulvinar projects to separate regions of extrastriate cortex. The posterior (IPp), medial (IPm) and central (IPc) nuclei of the inferior pulvinar are from Lin and Kaas. The subdivisions of visual cortex of the owl monkey are from Allman and Kaas. Areas VI (primary visual cortex), VII (secondary visual cortex), MT (middle temporal visual area), DL (dorsolateral visual area), and DM (dorsomedial visual area) each contain a topographic representation of the contralateral visual hemifield and have distinctive architectonic features. Areas PP (posterior parietal cortex) and DI (dorso-intermediate visual area) are visually responsive, but their topography has not been fully determined. The rostral dashed lines mark the extent of visually responsive cortex (V), which includes subdivisions not yet fully defined.

From: R.E. Weller and J. H. Kaas, "Connections of Visual Cortex in Primates," in C. N. Woolsey, Multiple Visual Areas, p. 126.

References

Allman, J.M., J.F. Baker, W.T. Newsome, and S.E. Petersen, "Visual topography and function: Cortical visual areas in the owl monkey." In Woolsey, C.N., *Cortical Sensory Organization.* Clifton, N.J.: The Humana Press Inc., 1981.

Amari, S. and M.A. Arbib (eds.), *Competition and Cooperation in Neural Nets*, Vol. 45 of *Lecture Notes in Biomathematics*, S. Levin (ed.). (Proceedings of the U.S.-Japan Joint Seminar, Kyoto, Japan, February 1982.) Berlin: Springer-Verlag, 1982.

Anderson, J.R. *Language, Memory, and Thought.* Hillsdale, NJ: L. Erlbaum, 1976.

Anderson, J.R. and G.H. Bower. *Human Associative Memory.* Washington, DC: V.H. Winston and Sons, 1972.

Ballard, D.H. and P. Coleman, "Cortical connections: structure and function," Working paper (TR), 1982.

Ballard, D.H., "Parameter networks," TR75, Computer Science Dept, U. Rochester, 1981; *Proc.*, 7th IJCAI, Vancouver, B.C., August 1981; revised May 1982.

Ballard, D.H. and C.M. Brown. *Computer Vision.* Prentice Hall, 1982.

Ballard, D.H. and O.A. Kimball, "Rigid body motion from depth and optical flow," TR70, Computer Science Dept., U. Rochester, November 1981.

Barlow, H.B., "Critical limiting factors in the design of the eye and visual cortex," *Proc. R. Soc. Lond., B 212*, 1-34, May 1981.

Barlow, H.B., "Reconstructing the visual image in space and time," *Nature*, Vol. 279, 17, May 1979.

Barrow, H.G. and J.M. Tenenbaum, "Recovering intrinsic scene characteristics from images." In Hanson, A.R. and E.M. Riseman (eds.), *Computer Vision Systems.* NY: Academic Press, 1978.

Blake, R., "The visual system of the cat," *Perception & Psychophysics*, 26 (6), 423-448, 1979.

Brady, M. (ed.). *Computer Vision.* Reprinted from the journal *Artificial Intelligence*, Vol. 17, Cambridge, MA: MIT Press, 1982.

Brewer, W.F. and J.C. Treyens, "Role of schemata in memory for places," *Cognitive Psychology*, 13, 207-230, 1981.

Brown, C.M., "Color vision and computer vision," TR108, Computer Science Dept., U. Rochester, June 1982.

Bruce, C., R. Desimone, and C.G. Gross, "Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque," *Journal of Neurophysiology*, Vol. 46, No. 2, August 1981.

Buser, P.A. and A. Roguel-Buser (eds.), *Cerebral Correlates of Conscious Experience.* Amsterdam: North Holland Publishing Co., 1978.

Carpenter, R.H.S. *Movements of the Eyes.* London: Pion Limited, 1977.

Collins, A.M. and E.F. Loftus, "A spreading-activation theory of semantic processing," *Psych Review, 82*, 407-429, November 1975.

Coren, S., "The interaction between eye movements and visual illusions." In Fisher, D.F., R.A. Monty, and J.W. Senders (eds.), *Eye Movements: Cognition and Visual Perception.* Hillsdale, N.J.: L. Erlbaum Associates, 1981.

Coren, S. and J.S. Girgus, "Illusions and constancies." In W. Epstein (ed.), *Stability and Constancy in Visual Perception.* New York: John Wiley & Sons, 255-284, 1977.

Cotman, C.W. (ed.). *Neuronal Plasticity.* NY: Raven Press, 1978.

Cowey, A., "Why are there so many visual areas?" In Schmitt, F.O., F.G. Warden, G. Adelman, and S.G. Dennis (eds.), *The Organization of the Cerebral Cortex.* Cambridge, Mass.: MIT Press, 1981.

Davidson, M.L., M.J. Fox, and A.O. Dick, "Effect of eye movements on backward masking and perceived location," *Perception and Psychophysics,* 14, No. 1, 110-116, 1973.

Epstein, W. (ed.), *Stability and Constancy in Visual Perception: Mechanisms and Processes.* New York: John Wiley & Sons, Inc., 1977.

Fahlman, S.E., "The Hashnet interconnection scheme," Computer Science Dept, Carnegie-Mellon U., June 1980.

Fahlman, S.E. *NETL, A System for Representing and Using Real Knowledge.* Boston, MA: MIT Press, 1979.

Feldman, J.A., "A connectionist model of visual memory." In G.E. Hinton and J.A. Anderson (eds.), *Parallel Models of Associative Memory.* Hillsdale, NJ: L. Erlbaum Associates, 1981.

Feldman, J.A., "Memory and change in connection networks," TR96, Computer Science Dept., U. Rochester, July 1981. To appear in *Biological Cybernetics.*

Feldman, J.A. and D.H. Ballard, "Connectionist models and their properties," to appear in *Cognitive Science,* 1982.

Ferster, D., "A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex," *J. Physiol.,* 311, 623-655, 1981.

Fisher, D.F., R.A. Monty, and J.W. Senders (eds.), *Eye Movements: Cognition and Visual Perception.* Hillsdale, NJ: L. Erlbaum Associates, 1981.

Gilchrist, A.L., "Perceived lightness depends on perceived spatial arrangement," *Science,* 185-187, 1977.

Graybiel, A.M. and D.M. Berson, "Families of related cortical areas in the extrastriate visual system: Summary of an hypothesis." In Woolsey, C.N., *Cortical Sensory Organization.* Clifton, N.J.: The Humana Press Inc., 1981.

Graybiel, A.M. and D.M. Berson, "On the relation between transthalamic and transcortical pathways in the visual system." In Schmitt, F.O., F.G. Warden, G. Adelman, and S.G. Dennis (eds.), *The Organization of the Cerebral Cortex.* Cambridge, Mass: MIT Press, 1981.

Gross, C.G., C.J. Bruce, R. Desimone, J. Fleming and R. Gattass, "Cortical visual areas of the temporal lobe: three areas in the Macaque." In Woolsey, C.N., *Cortical Sensory Organization.* Clifton, N.J.: The Humana Press Inc., 1981.

Haber, R.N., "Stimulus information and processing mechanisms in visual space perception." In Rosenfeld and Beck (in preparation), 1983.

Haber, R.N. (ed.), *Contemporary Theory and Research in Visual Perception.* New York: Holt, Rinehart and Winston, Inc., 1968.

Hanson, A.R. and E.M. Riseman (eds.), *Computer Vision Systems.* NY: Academic Press, 1978.

Hebb, D.O. *The Organization of Behavior.* NY: Wiley, 1949.

Hillis, W.D., "The connection machine (Computer architecture for the new wave)," AI Memo 646, M.I.T., September 1981.

Hinton, G.E., "Shape representation in parallel systems," *Proc.*, 7th IJCAI, 1088-1096, Vancouver, B.C., August 1981.

Hinton, G.E., "The role of spatial working memory in shape perception," *Proc.*, Cognitive Science Conf., 56-60, Berkeley, CA, August 1981.

Hinton, G.E. and J.A. Anderson (eds.), *Parallel Models of Associative Memory.* Hillsdale, NJ: L. Erlbaum Associates, 1981.

Hintzman, D.L., C.S. O'Dell, and D.R. Arndt, "Orientation in cognitive maps," *Cognitive Psychology*, 13, 149-206, 1981.

Hochberg, J., "Perception of successive views," *Science*, 1978.

Hochberg, J. *Perception.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1964.

Horn, G., G. Stechler, and R.M. Hill, "Receptive fields of units in the visual cortex of the cat in the presence and absence of bodily tilt," *Experimental Brain Research*, 15, 113-132, 1972.

Howard, I.P. *Human Visual Orientation.* New York: John Wiley & Sons, 1982.

Hrechanyk, L.M. and D.H. Ballard, "A connectionist model of form perception," *Proc.*, IEEE Workshop on Computer Vision, Rindge, N.H., 44-51, Aug. 1982.

Johansson, G., "Spatial constancy and motion in visual perception." In Epstein, W. (ed.), *Stability and Constancy in Visual Perception: Mechanisms and Processes.* New York: John Wiley & Sons, Inc., 1977.

Jonides, J., D.E. Irwin, and S. Yantis, "Integrating visual information from successive fixations," *Science*, Vol. 215, 192-194, January 1982.

Jusczyk, P.W. and R.M. Klein (eds.), *The Nature of Thought: Essays in Honour of D.O. Hebb.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Kinsbourne, M. and R.E. Hicks, "Functional cerebral space: A model for overflow, transfer and interference effects in human performance: A tutorial review." In J. Requin (ed.), *Attention and Performance 7.* New Jersey: Lawrence Erlbaum Associates, 1979.

Kosslyn, S.M. *Images and Mind.* Cambridge, MA: Harvard U. Press, 1980.

Kuipers, B.J., "Modelling spatial knowledge," *Cognitive Sci.*, 2: 129-153, 1973.

Kuffler, S. W. and J.G. Nicholls, *From Neuron to Brain: A Cellular Approach to the Function of the Nervous System.* Sunderland, MA: Sinauer Associates, Inc., 1976.

Lynch, J.C., "The functional organization of posterior parietal association cortex," *The Behavioral and Brain Sciences*, 3, 485-534, 1980.

Lynch, K. *The Image of the City.* Cambridge, MA: The M.I.T. Press, 1960.

Macko, K.A., et al., "Mapping the primate visual system with [2-^{14}C] deoxyglucose," *Science*, Vol. 218, No. 4570, 394-396, October 1982.

Marr, D.C. and T. Poggio, "Cooperative computation of stereo disparity," *Science, 194*, 283-287, 1976.

Marr, D.C. *Vision.* San Francisco, Ca: W.H. Freeman and Co., 1982.

Mays, L.E. and D.L. Sparks, "Saccades are spatially, not retinocentrically, coded," *Science*, Vol. 208, June 1980.

McClelland, J.L. and D.E. Rumelhart, "An interactive activation model of the effect of context in perception: Part 1," *Psychological Review*, 1981.

McDermott, D. "Spatial inferences with ground, metric formulas on simple objects," Yale University, Department of Computer Science, Research Report #173, January 1980.

Menzel, E.W., "Chimpanzee spatial memory organization," *Science*, 182, 943 945, 30 November 1973.

Mesulam, M.M., "A cortical network for directed attention and unilateral neglect," *Annals of Neurology*, Vol. 10, No. 4, October 1981.

Minsky, M. and S. Papert. *Perceptrons.* Cambridge, MA: The MIT Press, 1972.

Montero, V.M., "Topography of the cortico-cortical connections from the striate cortex in the cat," *Brain, Behavior and Evolution*, 18, 194-218, 1981.

Morgan, M.J., "How pursuit eye movements can convert temporal into spatial information." In Fisher, D.F., R.A. Monty, and J.W. Senders (eds.), *Eye Movements: Cognition and Visual Perception.* Hillsdale, NJ: L. Erlbaum Associates, 1981.

Norman, D.A., "A psychologist views human processing: Human errors and other phenomena suggest processing mechanisms," *Proc.*, IJCAI, 1097-1101, Vancouver, B.C., August 1981.

Palmer, S.E., "Transformational structure and perceptual organization," *Proc.*, Third Annual Meeting, Cognitive Science Society, Berkeley, CA., August 1981.

Phillips, W.A. and D.F.M. Christie, "Components of visual memory," *Quarterly J Experimental Psychology, 29*, 117-134, 1977.

Piaget, J. and B. Inhelder. *The Child's Conception of Space.* New York: Norton, 1967.

Pinker, S. and R.A. Finke, "Emergent two-dimensional patterns in images rotated in depth," *Journal of Experimental Psychology: Human Perception and Performance*, 6, No. 2, 244-264, 1980.

Posner, M.I. *Chronometric Explorations of Mind.* Hillsdale, NJ: L. Erlbaum Associates, 1978.

Robinson, D.L., M.E. Goldberg, and G.B. Stanton, "Parietal association cortex in the primate: Sensory mechanisms and behavioral modulations," *Journal of Neurophysiology*, 41, No. 4, July 1978.

Roland, P.E., E. Skinhoj, N.A. Lassen, and B. Larsen, "Different cortical areas in man in organization of voluntary movements in extrapersonal space," *Journal of Neurophysiology*, 43, No. 1, January 1980.

Sabbah, D., "Design of a highly parallel visual recognition system," *Proc.*, 7th IJCAI, Vancouver, B.C., August 1981.

Sakata, H., H. Shibutani, and K. Kawano, "Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey, "*Journal of Neurophysiology*, 43, No. 6, 1654-1672, June 1980.

Schlag, J., M. Schlag-Rey, C.K. Peck, and J.P. Joseph, "Visual responses of thalamic neurons depending on the direction of gaze and the position of targets in space," *Experimental Brain Research*, 40, 170-184, 1980.

Schmitt, F.O., F.G. Warden, G. Adelman, and S.G. Dennis (eds.). *The Organization of the Cerebral Cortex*. Cambridge, Mass: MIT Press, 1981.

Shebilske, W., "Visual direction illusions in everyday situations: implications for sensorimotor and ecological theories." In Fisher, D.F., R.A. Monty, and J.W. Senders (eds.), *Eye Movements: Cognition and Visual Perception*. Hillsdale, NJ: L. Erlbaum Associates, 1981.

Small, S., G. Cottrell, and L. Shastri, "Toward connectionist parsing," *Proc.*, National Conference of American Association of Artificial Intelligence, Pittsburgh, PA., August 1982.

Smith, E.E., E.J. Shoben, and L.J. Rips, "Structure and process in semantic memory: A featural model for semantic decisions," *Psychological Review*, 81, 3, 214-241, 1974.

Stent, G.S., "A physiological mechanism for Hebb's postulate of learning," *Proc.*, National Academy of Science USA, 70, 4, 997-1001, April 1973.

Sutton, R.S. and A.G. Barto," Toward a modern theory of adaptive networks: Expectation and prediction," *Psychological Review*, 88, 2, 135-170, 1981.

Tomko, D.L., N.M. Barbaro, and F.N. Ali, "Effects of body tilt on receptive field orientation of simple visual cortical neurons in unanesthetized cats," *Experimental Brain Research*, 43, 309-314, 1981.

Torioka, T., "Pattern separability in a random neural net with inhibitory connections," *Biological Cybernetics*, 34, 53-62, 1979.

Treisman, A., "The role of attention in object perception," *Proc.*, The Royal Society International Symposium on Physical and Biological Processing of Images, London, Sept. 1982.

Tréisman, A.M. and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology, 12*, 97-136, 1980.

Turvey, M.T., "Contrasting orientations to the theory of visual information processing," *Psychological Review, 84*, 67-88, 1977.

Ullman, S., "Filling-in-the gaps: The shape of subjective contours and a model for their generation," *Biological Cybernetics*, 25, 1-6, 1976.

Van Essen, D.C., J.H.R. Maunsell and J.L. Bixby, "Organization of extrastriate visual areas in the macaque monkey." In Woolsey, J., *Cortical Sensory Organization*. Clifton, NJ: The Humana Press Inc., 1981.

Volkmann, F.C. "Saccadic suppression: a brief review." In R.A. Monty and J.W. Senders (eds.), *Eye Movements and Psychological Processes*. Hillsdale, NJ: L. Erlbaum Associates, 73-84, 1976.

Weisstein, N. and W. Maguire, "Computing the next step: Psychological measures of representation and interpretation." In Hanson, A.R. and E.M. Riseman (eds.), *Computer Vision Systems*. New York: Academic Press, 1978.

Welch, R.B. *Perceptual Modification: Adapting to Altered Sensory Environments.* New York: Academic Press, 1978.

Wickelgren, W.A., "Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system," *Psychological Review, 86*, 1, 44-60, 1979.

Wilson, H.R., and J.R. Bergen, "A four mechanism model for spatial vision," *Vision Research*, 19, 19-32, 1979.

Wurtz, R.H. and J.E. Albano, "Visual-motor function of the primate superior colliculus," *Ann Rev Neurosci, 3*, 189-226, 1980.

Zeki, S., "The representation of colours in the cerebral cortex," *Nature, 284*, April 1980.