Time, Space and Form in Vision

Jerome A. Feldman[1,2]

TR-88-011

November, 1988

Also available as:
TR244
November 1988
Department of Computer Science
University of Rochester
Rochester, New York 14627

ABSTRACT

The prodigious spatial capabilities of the primate visual system are even more remarkable when temporal considerations are taken into account. Recent advances in neurophysiology, phychophysics and computer vision provide significant constraints on how the system could work. This paper presents a fairly detailed connectionist computational model of how the perception and recognition of objects is carried out by primate brains. The model is claimed to be functionally adequate and to satisfy all the constraints established by the various disciplines. One key notion introduced is a multi-input, multi-output network for inverting spatio-temporal cues. The central construct in intermediate level motion vision is taken to be the *trajectory* and these are used in recognition of dynamic situations called *scenarios*. The entire development is an extension of the author's 1985 Four Frames model, which required relatively little modification to accommodate temporal change (eventually).

[1]International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California 94704.
[2]Department of Computer Science, University of California at Berkeley, Berkeley, CA 94720.

This work was supported in part by ONR grant No. N00014-84-0655, and in part by NSF Coordinated Experimental Research grant No. DCR-8320136.

I. Introduction

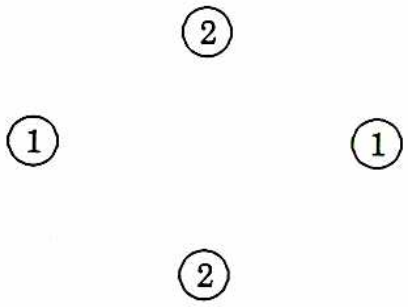
The exquisite spatial capabilities of the primate visual system are even more remarkable when the adverse temporal conditions are taken into account. Changes of ecological importance can be significantly faster than the integration times of photoreceptors or the signalling ability of optic nerve fibers. The remarkable dynamic range ($\sim 10^{10}$) of the system is achieved by fairly rapid adaptation [Haber & Hershenson, p. 53] and a variety of large and small eye movements must also be taken into account. This paper is an attempt to outline an overview of temporal change in the understanding of visual scenes. It is a direct continuation of an effort begun about a decade ago, the early results of which were presented and criticized in [Feldman 1985].

[Feldman 1985] presents an explicit computational model of spatial phenomena in vision that was claimed to be consistent with all relevant findings (and to be the only such model extant). Aside from one controversial issue to be discussed in Section V, the model has not been seriously challenged. (Nor, it must be admitted, taken seriously in any other way). But that entire enterprise explicitly suppressed all questions of temporal change. Much of my effort in the subsequent years has been devoted to getting some understanding of the temporal properties of the visual system and how it deals with change. This paper is one of a set of reports that are in varying stages of completion and references to the others will be in braces. One paper by Brian Madden {Madden 1988a} describes a wide variety of studies in apparent motion. Another {Madden 1988b} contains a review of the literature on space and time plus some rigorous psychophysical studies on estimation from apparent motion. Tom Olson's thesis {Olson 1988} deals with the problem of computing trajectories from visual input while Nigel Goddard's {Goddard 1988} is concerned with using trajectory data for recognition, especially of Johannsen-type moving light displays [Johannsen 1973]. The current verison of any of these can be requested through the Computer Science Department at Rochester.

The normal functioning of the visual system is so robust that it is often necessary to use abnormal conditions to study it. One major source of constraints for this paper comes from studies of apparent motion. Example 1 will help illustrate some of the issues of concern:

If a subject is first shown the two small circles labelled ① and, after a delay of 50-500 milliseconds, the two circles labelled ②, a strong perception arises that two dots simultaneously moved at an appropriate speed. The distance between the dots can be several degrees, far beyond the range of receptive fields of neurons in the retina or primary visual cortex. The continuous spatio-temporal change (*slip*) which is typically the basic motion cue is totally absent. Also, it is interesting that the motion path (*trajectory*) must be retroactively determined by the second flash. Most importantly, the display of Example 1 is ambiguous -- the dots can be seen as moving clockwise or counter-clockwise. A remarkable fact is that, under these and most other conditions, the system always chooses consistent interpretations for all the

Example 1:



moving dots. The problem of ambiguous correspondence or *match* targets also holds at small separations; any difference or Reichardt type theory of change perception must address this problem and we discuss this further in Section III.

When the dots in Example 1 are far apart, there are several influences on which of the competing motions is perceived {Madden 1988a}. Slight biases in the spatial and temporal differences are effective, as are matching the shapes, colors or contrasts of two dots. A subject can almost always choose which motion to see. Adding even one (17 m.sec) frame of slip will bias the choice as will priming from previous examples, neighboring displays, etc. Adding a static, low-contrast (blur) path between two dots is compelling [Shepard & Zare 1982]. Furthermore, {Madden 1988b} shows that subjects are fairly good (accuracy ~10%) at estimating speed and direction from two-flash displays. All of these effects are representative of general properties of normal motion and change processing and will reoccur several times in the sequel.

The purpose of this paper is to model these and related phenomena in a way that makes biological and computational sense. Example 1 is obviously highly artificial, but such ambiguities arise in any sitatuion with similar-looking moving objects. My starting assumption is that the visual system evolved to construct plausible real world scenarios that make sense out of the ongoing spatio-temporal flux in the context of the animal's current internal state. This is a direct extension of my atemporal four-frames model [Feldman 1985] and a brief reprise of that model seems appropriate.

Four Frames Reprise

The basic idea is that vision is carried out by a collection of interacting networks grouped into four distinct representational frames of reference, (cf. Figure 1). The units in all these networks should be thought of as abstract neurons, computing activity levels and communicating by simple codes. 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 has been 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 what the first foveal-periphery frame is actually 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 two frames. Of course, neither of these frames is like a photographic image of the world. Light striking the retina is already transformed, and the layers of the retina, the thalamus, and the visual cortex all compute complex functions. The crucial difference between these two frames is that the first one is totally updated with each saccade, while the second is not. The model also assumes that the first (*retinotopic*) frame computes proximal stimulus features and the second captures distal (constancy, intrinsic) features in addition to being stable. The latter is therefore called the *stable feature frame*. An important aspect of the model is the assumption that phenomenal perception as well as recognition is based on constancy, not retinotopic features.

The third and fourth representational frames are both multi-modal and are thus unlikely to be the same as the first two. The third representation is not primarily geometrical and will be described in the next paragraph. The fourth, or *environmental frame*, 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 that gets mapped, by *situation links*, to the current situation and the observer's place in it. Not treated in either the 1985 paper or this one are physical actions and the coordinate frames used for them.

The third representational frame 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 propositional (relational) form, modeled in this case by a kind of semantic network. One class of knowledge encoded will be the visual appearance of objects encoded as a collection of relationships among primitive parts. These descriptions have much of the character of Minsky's conceptual frames [1975] and of the object-centered frames of, for example, Ballard [1984] and Hinton [1981]. Since the other three representations are geometrically organized, the collection of semantic knowledge will be referred to as the *world knowledge formulary*, to emphasize its nature as a collection of conceptual relations. The formulary carries much of the burden for integrating information from the other three frames and is far from adequately worked out [Shastri 1988]. But all that is needed for now is the notion that the network representation is likely to be quite different from that of the retinotopic, the feature, or the environmental frame. All of this suggests that even a provisional model of vision and space will require at least four representational frames.

The central problem of vision is linking visual-feature information to 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*, by analogy to looking up something in an index. Obviously enough, it is more effective to index with invariant, real-world features than with their retinal manifestations and facilitating this is the primary function of the stable feature frame. Recognition of an object or situation is modeled as a mutually reinforcing coalition of active nodes in the world knowledge frame. The mutual excitation of feature and model networks also involves top-down, *context*, links from visual elements to the feature units that are appropriate (cf. Figure 1). We note in passing that the indexing process involves "de-spacing" the feature information; there could not be separate recognition networks for an object in each position in space. This will be discussed in Section VII in connection with the what-where distinction in visual processing.

Figure 1 about here.

The first notion of appearance models mentioned above 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. One can probably name an object that is an approximately 1.5 inch white sphere and 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 is not even the right kind of visual information. Our present way of handling the appearance models for complex objects and situations is taken from current AI practice. It is assumed 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 of the relationships among them (cf. Figure 2). 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 hierarchical network is all that is needed at the moment [Cooper 1988].

The basic idea is that each visual element of a complex object is represented by a node that corresponds to a particular set of feature values as computed in the feature frame. Since indexing from features to elements occurs in parallel, there will usually be several simultaneously active element 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 the details of complex object representations are considered, a number of difficult technical problems arise. This is discussed in detail in Hrechanyk and Ballard [1982] and in Section VI, and a discussion in outline will suffice, based on the example of representing the visual appearance of horses. Recall that the world knowledge formulary visual-appearance models are far from complete -- they are more like verbal descriptions 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. One must also account for the fact 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 2, involves instance nodes, separate sub-networks for different views and cross-referenced structural descriptions. The prototype horse has a general hierarchical description in which, e.g., the trunk is composed of a body, legs, and a tail. What visual elements 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 elements. As always, there is assumed to be mutual inhibition among competing object descriptions and view nodes. There is a good deal of ongoing work on this kind of recognition network [Cooper & Hollbach 1987; Plaut 1984; Cooper 1988] and it does appear to be computationally reasonable. A major goal of this paper is to extend these ideas to recognition of *motions*, such as a horse's canter. This involves reworking all levels of the four-frames model and confronting several important issues elided in the atemporal treatment.

Figure 2 about here

Temporal Preliminaries

One reason that it has been difficult to isolate the temporal properties of the visual system is that the internal time scale of the computing elements is of the same order as the events they are trying to compute. Nature has been constrained to elements with millisecond operation times and every aspect of the system exhibits signs of this constraint. It is true that there are special adaptations to detect binaural differences that translate to a much smaller (microsecond) time scale [Knudsen *et al.* 1987], but these are not well understood and there is no indication that they are used in normal vision. The rise time of photo-receptor potentials, the transmission times of axons and the firing rate of neurons all lie in the range of a few to a few tens of milliseconds. The performance of the visual system in resolving time differences, speeds, flicker rates, etc. fits nicely into the same range of times, but we need to understand how it is done. Pulfrich illusions caused by reducing the light to one eye show that even the notion of relative time of visual events can be easily confounded. An additional conceptual problem arises because much of the information in the system is transmitted by a temporal code (spike rate) which requires more time than the events it is describing. It does not seem likely that we will be able to treat temporal change as just another visual property like color or spatial scale. The fundamental complexity of temporal issues in form vision extends through all conceptual and anatomical levels. A large body of evidence suggests that there are two parallel pathways extending through many anatomical levels and

characterized, at least in part, by different temporal characteristics [Maunsell 1987], and Figure 6. Taking temporal change as an organizing principle provides a new perspective on several classical vision problems. Appendix A is a list of the major temporal questions addressed in this paper along with the sections in which they are treated. As in [Feldman 1985], it is claimed that no significant issue has been overlooked and that the answers provided are consistent with each other and with the relevant behavioral, biological and computational findings. Also, unfortunately, it still appears that there is no alternative model in the literature.

Another preliminary observation is that temporal change in visual information has many possible real-world causes and is detected by several mechanisms of the visual system. All of these interact in complex ways and it is difficult to isolate one mechanism experimentally or theoretically. Figure 3 presents the main inputs and outputs of the visual change processing system and is the skeleton for much of the rest of the paper. The inputs (above) represent various ways that the visual system obtains information pertaining to temporal change. The lower half depicts the kinds of real-world events to which the system can attribute the changes. These include a perception of self-motion, the movement of articulated objects within a scene and non-rigid shape changes. We are also able to detect other sources of image change such as changes in illumination. One important point is that there is no simple relation between the kind of real-world change and its manifestations. The situation is exactly analogous to the problems in spatial vision where the brightness at a point is a joint function of illumination, distance, albedo and orientation. The visual system apparently solves these inverse problems by best-fit in a parameter space embodied in neural networks. Since the solution is normally rapid and robust, the networks must embody much of the solution in their structure.

Figure 3 about here

The main inputs to the change processing network come from three sources: Pursuit eye movements, local (short-range) slip detection, and the matching (correspondence) of features displaced in space and time. In addition, estimates of depth (from stereopsis, etc.) and top-down contextual expectations play a central role in the interpretation of visual change. The first half of the paper (Sections II-IV) is largely concerned with what is known about the inputs and outputs of Figure 3 and their interactions. But, as in [Feldman 1985], constancy feature calculations form just the base of our concerns. The remainder of the paper explores how the change processing network of Figure 3 could fit into the visual system modelled in Figure 1. Basically, the motion and change network of Figure 3 is an elaboration of the "motion" stable feature frame of Figure 1, as depicted in Figure 9. It is the computational detail of this elaboration that the paper addresses. One question concerns how the invariant temporal computations might interact with the other

invariant features computed in the stable feature frame and some related questions involving eye-movements and other changes (Section V).

The indexing process that links the feature frame with world knowledge retains its central role. A specific motion primitive, the *trajectory*, is hypothesized as the key link between features and objects. The way in which trajectories are computed and used is the basis for the last part of the paper. The static "situations" of the earlier model are extended to multi-temporal "scenarios" and many of the previous mechanisms are extended (Section VI). Finally, an attempt is made to make computational sense of the hypothesized "what" and "where" dichotomy between the ventral and dorsal branches of visual cortex (Section VII).

II. Basic Temporal Issues

In addition to their central role in change processing, temporal considerations are at the heart of many other puzzles about the visual system. Some of these were mentioned in the introduction and will recur throughout the paper. Appendix A lists the issues treated in various sections. This section concentrates on some basic temporal properties of the visual system that underlie all the later discussions. These include inseparability at the receptor, retinal smear, masking and meta-contrast, adaptation and fading and the supposed benefits of moving images. All of these issues have been controversial and none is fully resolved, but a coherent picture is beginning to emerge.

The paper will focus on temporal issues in the range stretching from milliseconds to minutes, about seven orders of magnitude. There are important issues with faster and slower time courses, but they appear to be separable. As in the earlier paper, I combine human psychological data with physiologial results from other primates on the assumption that human physiology is sufficiently similar for my purposes. For concreteness, the discussion will center on vision in the normal daylight range (10^3 to 10^4 trolands) and thus be more concerned with cones, photopic psychophysics, etc..

Receptor level dynamics

One anchor point for our discussion is the time course of interaction in the visual receptors. Figure 4 (from Baylor [1987]) shows the time course of response of macaque receptors to short pulses of light of varying intensity. There are several important pieces of information in this figure. First notice that, over a broad range of intensities, the cones take about 50 milliseconds to reach their peak response and about the same amount of time to decay to zero. The time to return to the base state is about 200 milliseconds, in good correspondence with the psychophysical finding that the peak flicker sensitivity of human cone vision is about 5 hertz. However, people can detect onset difference of only 2-4 milliseconds for flashes separated by a few minutes of arc [Westheimer & McKee 1981] and thus must detect some transient properties of the curves in Figure 4. A number of the basic properties of temporal vision are understandable from the data of Figure 4. For example, at normal intensities the system appears to be unable to distinguish length of stimulus durations less than 150 milliseconds and essentially counts total photons [Haber & Hershenson, p. 56]. We will have to deal later with "inseparability at the receptor" -- the fact that individual receptors have no way to distinguish intensity changes caused by moving objects from those caused by variations in light or by motions of the eye. Spatio-temporal inseparability remains an issue at high levels of the visual system [Fleet *et al.* 1985].

Figure 4 about here

Another way to view Figure 4 is to think of the cone curves as describing a temporal smoothing process at the first stage of visual processing. This appears to have the same beneficial anti-aliasing effect in time as the point-spread function of the lens has for spatial vision [Williams 1986]. In both cases, the sampling induced by physical resolution limits (in space or time) is nicely matched by prior smoothing operations. In the temporal case, the physical bound is the limited rate at which spiking cells (starting with retinal ganglion cells) can transmit information about changes. The high spatial resolution subsystem responsible for conveying detailed contrast and form (and color) information is constrained to relatively slow temporal changes by both receptor and ganglion cell dynamics. An additional factor is that the smaller, more precise, ganglion cells integrate over less area and take more time to accumulate an adequate signal. As the details continue to be worked out [*Trends in NeuroScience*, special issue], it becomes ever clearer how this system, corresponding to physiological P (X in cat) ganglion cells and anatomical β-cells does its work. An important part of the puzzle that has not yet been integrated is adaptation. It turns out that adaptation to different light levels is known psychophysically to have a temporal and spatial range that overlaps the signals themselves [Hayhoe *et al.* in press]. There are clearly adaptations in the receptors and much of the local circuity of the retina is also concerned with adaptation [*Trends in NeuroScience*, special issue]. Again this processing must be pre-digitization and is being modelled as such in hardware [Hutchinson *et al.* 1988]. One (non-temporal) point that does not seem to be well interpreted in the standard accounts is that there is also good information on absolute intensity computed and sent on [Barlow 1981]. Madden {1988b} reviews many basic facts on adaptation and temporal vision in a framework compatible with this paper.

A related issue concerns the role of retinal smear in vision. Even at a moderate speed of 10°/sec. a moving object will cover about a degree in the approximately 100 millisecond sensor integration time, passing over many receptors. Why isn't there a smeared image? It turns out that there is retinal image smear as Figure 5 (after [Burr *et al.* 1986]) helps show. Assuming a conservative 125 millisecond fixation, this figure shows how blurred the images of walking people would be. In fact, people expect a certain amount of retinal blur to accompany motion and animation engineers have found that they need to add it to rapid motion sequences or people complain [Dippé & Wold 1985]. Furthermore, Shepard and Zare [1982] have shown that a blur path shown in an apparent motion sequence will strongly bias perception of motion to be along that path. The blur caused by saccades is also normally suppressed [Volkman 1986]. Again, we don't notice the blur path for the same reason we don't notice retinal size or luminance; "noticing" is at higher processing levels. The parcelization of various change cues back into presumed real-world causes is the main topic of Section IV of this paper.

Figure 5 about here

The curves in Figure 4 also suggest how the visual system can discriminate velocities much better than the 100 millisecond cone integration time would suggest. All of the cone response curves show a very steep rise and the retina could have mechanisms for detecting these transients. The mathematical form of these transients is some kind of temporal derivative. Notice that any such mechanism would have to be analog and not based on spike frequencies. There are two related reasons for this. For a frequency code to be generated, some receptor cell would have to have first integrated the signal and generated the code; this is just the problem that we are trying to overcome. Furthermore, the integration of this receptor would remove the high temporal frequency information needed for the calculation. As has been understood for some time, there is a sub-network in the vertebrate retina which seems primarily concerned with reporting transients or temporal change. This is identified with the physiological M (Y in cat) and anatomical α ganglion cells and is also being worked out. Velocity information could be computed in the retina and is in many animals, but primates have essentially no direction selective cells before visual cortex. A separate system with high temporal resolution is needed to transmit the required information (over low bandwidth channels) to the areas where velocity is computed. Koch *et al.* in the same *Trends in NeuroScience issue* [May 1986] present evidence that the unit response sharpens with succeeding retinal processing and present detailed biophysical models of how this might come about.

So we see that, at the retinal level, there is a division of the information into a slower and more precise (along several dimensions) stream and a second stream that is cruder, but responds rapidly to change. The complete story is much more complex involving adaptation and a variety of signal processing mechanisms, but the temporal division is essential for all further processing. There is now clear evidence that a division related to the M-P split continues through the visual system, with many points of interaction between the pathways. Figure 6 depicts the basic anatomy of the so-called transient and sustained systems. Psychophysical experiments also strongly suggest two broad temporal channels [Thompson; Anderson & Burr 1985]. As we have seen, the more precise (in space and intensity resolution) P system does not have adequate temporal resolution and is beautifully complemented by the rapidly responding, but cruder M system at the retinal level. But the inadequate signalling speed of neurons continues throughout the visual system; this might be part of the reason why parallel pathways have evolved. In several sections of this paper, suggestions will be made on some of the ways that a slow accurate system might interact with a more responsive one to yield improved performance at different tasks. For example, in the next section, we indicate how local motion (slip) calculation could benefit from the two pathways. There are also several other ways of dividing the visual pathway into channels; one important way is by size or spatial frequency. Although it will not be mentioned explicitly, one should think of the processes described here as occurring at several spatial scales.

Figure 6 about here

Masking and Stabilization

One temporal interaction that is relevant here is "transient-on-sustained inhibition" [Breitmeyer 1984]. It makes computational sense that the static form processing subsystem be inhibited during periods when the image is changing too rapidly, such as during a saccade. Such situations automatically induce blurred images and thus lower contrast and less response. But it is also known that there are specific inhibitory effects of the transient on the sustained channels [Volkman 1986]. There have been attempts to explain much of vision in this way [Breitmeyer 1984], but the evidence suggests a much more complex set of interactions. Even saccadic suppression is known to have an efferent component which can raise thresholds some 30 milliseconds before a saccade [Carpenter, p. 248].

Some of the most complex issues in temporal vision concern masking and related phenomena such as meta-contrast. From our connectionist perspective, at least some coherence merges. Signals in highly interconnected feedback networks will combine in complex ways. Stimuli that are close enough together in time are merged in processing. Some masking is just direct interference early in the channels and thus in our retinotopic frame, but other kinds must occur at higher levels. It is also important to notice a kind of inversion in terminology from phenomena to mechanism. Thus "backward masking" (interference with an earlier signal by a later one) is forward action in the network. The famous U-shaped curves that show maximum interference at 50-100 millisecond (SOA) delays suggest that some critical interactions for recognition have vulnerable interactions of that scale. Meta-contrast, i.e., backward masking by a surrounding stimulus, suggests that some of this interference is related to "filling in" calculations [Stoper & Mansfield 1978]. These interactions are assumed to be in the stable feature frame and constitute some of the evidence for it [Davidson *et al.* 1973].

Consider first the well-known fact that images that are stabilized to appear at the same place on the retina eventually fade away. This has often been taken to imply that drift, tremor and micro-saccades are necessary for normal vision. The most important point here is a purely temporal one -- the fading of stabilized images is many times slower than the receptor integration time and the rate of eye movements [Haber & Hershenson, p. 153]. The adaptation explanation of fading, which seems most likely, is consistent with the observed results. The reports that parts of the image disappear and reappear in coherent groups makes excellent sense in a model where phenomenal perception lies at the constancy, or stable feature frame level. This also suggests that there will be significant top-down contextual influences on how images are seen to fade, but this does not seem to have been tested.

A related claim in the literature is that the existence of motion and size sensitive receptors (a.k.a. frequency selective space-time filters) means that form vision is not harmed by retinal slip. This claim is immediately challenged by the existence of an enormous amount of mechanism dedicated to tracking moving objects to keep their images stable on the retina (Section IV). Motion sensitive mechanisms do not, and could not, have the spatial selectivity of static ones. It is true that detection and very simple discrimination tasks can be aided by motions up to a few degrees per second, but one would expect this of a two-channel system.

Example 2:

A frequently-cited argument for the benefits (or at least harmlessness) of retinal slip comes from the vernier acuity task depicted in Example 2. As is well known, people can distinguish left or right offset of two lines down to a separation (~15" of arc) rather less than a cone width. It turns out that a retinal velocity of up to about 3°/sec does not increase the least perceivable difference [Westheimer & McKee 1977]. This suggests to some people that the visual system retains its full resolving power for drifting images. A much more plausible explanation is that the vernier task is not testing fine resolution but just the discrimination between two distinct patterns. As suggested by the ovals in Example 2, many detectors (filters) will respond differently to the opposing vernier stimuli [Morgan & Watt 1984, Madden 1985]. There are other tasks, such as width discrimination, that are badly disrupted by slow drift [Welch & McKee 1985]. A related finding is that people confound spatial and temporal differences in the vernier range [Burr, Ross & Morrone 1986; Morgan & Watt 1982].

One also reads that miniature eye movements (drift, tremor, and micro-saccades) are ingenious adaptations to prevent fading. No one has pushed hard for tremor, but both drift [Kowler & Steinman 1980] and small saccades [Ditchburn 1980] are proposed as the basic mechanism to preserve vision. A much more likely analysis is presented by Carpenter [1977, p. 262]. Drift and tremor are natural properties of noisy neuro-muscular control systems and in natural situations there is additional destabilization due to head movements, etc. This provides more than enough change to keep the adaptation mechanisms in check. Small saccades, like large ones, re-direct foveal imaging. The question is not why the system introduces miniature movements, but how it overcomes them [Steinman 1986].

We do not know how the system accommodates miniature eye movements, but there is a likely possibility. Notice that tremor, which can be quite pronounced in some conditions, could not be "taken into account." But if we assume that the finest resolution visual mechanisms are primarily concerned with relative information (such as texture), then small motions will usually leave the information intact. To the extent that drift and micro saccades are programmed, they could be mapped as described for large movements in Section V. Assuming, as always, that subjective perception lies at the stable feature level, we would not notice miniature eye movements. In the autokinetic effect [Carpenter 1977, p. 272], a small isolated spot does seem to undergo random motions. The reason for this, in the current model, is the absence of contextual information (of the stable feature frame) to integrate the small fluctuations.

The main temporal issue in vision is the perception of motion and change. The next two sections consider this in some detail. Section three concentrates on local indications of temporal change (retinal slip) and mechanisms for processing this information. Section four expands the discussion to use other cues for motion and change, primarily proprioceptive cues and a long-range matching process. The various cues for change are postulated to combine in one of the stable feature frames to yield a distal or causal interpretation of the situation.

III. Retinal slip and the short-range mechanism

Any animal that moves under visual guidance relies heavily on the processing of continuously changing visual information. Within the timing limitations of the physiology, even very primitive animals approximate continuous, analog processing of these signals. For simple creatures such as the fly, this optical flow of information is the fundamental source of guidance. Primates, which are our main concern, also use pursuit eye movements and a long-range match mechanism but still depend crucially on low-level, local, continuous change detection.

A large fraction of current visual change research in psychophysics, physiology and computation has been devoted to understanding and modelling the processing of short-range continuous change. There have been several recent journal issues and surveys of this work [Nakayama, Hildreth & Koch 1987] which I will not attempt to further summarize. The issues that concern us are what the short-range or slip mechanism tells us about the temporal properties of the visual system and how it fits into the general change-processing network of Figure 3.

There is now ample evidence that there are at least two distinct mechanisms for visual change processing in primates [Baker & Braddick 1985]. We will use the term *slip* mechanism as synonymous with Braddick's term "short-range" and will also sometimes refer to Braddick's "long-range" motion as being processed by a *match* mechanism. The terminology is chosen to reflect the operation of the two mechanisms and would hopefully help reduce the confusion in the literature. Retinal slip is the continuous spatio-temporal change of light at the eye as reflected in continuous changes in membrane potentials of photoreceptors. The information derived from this continuous change is what is referred to as slip information in this paper. It is sometimes a cue for motion, but as we will see, it can also indicate tracking error, lighting change or a static object viewed with a moving eye. Our visual systems can also measure motion and other change quite accurately {Madden 1988b} in the absence of any slip information. This ability, most obvious in discontinuous or apparent motion, relies upon *matching* objects in two or more temporally separated frames. The double dissociation of the slip and match mechanisms comes from a wide range of experiments and follows a pattern already familiar from the last section.

The slip mechanism is distinguished from the match mechanism by many properties that one would expect in early vision. The slip subsystem is monocular, it is adaptable and yields motion after-effects [Baker & Braddick 1985]. It is relatively insensitive to stimulus features except for contrast reversal [Anstis 1980]. It is called short-range because it is only effective for small dots to a separation (d_{max}) of about 15 arc-minutes in the fovea. At an eccentricity of 10°, d_{max} approaches 100 minutes of arc, but this is still an order of magnitude less than the range of the long-range match mechanism. These results are independent of timing as long as the inter-stimulus interval is in the range of 5-80 m.seconds and the flashes are short, a millisecond or less. All of this is quite consistent with a story that slip could be

detected by the stimulation of nearby cones at slight delays. There is evidence that this happens, e.g. in rabbits [Barlow & Levick 1965], but things cannot be so simple in primates. This is because cells with significant directional selectivity do not appear in primates until visual cortex, two major steps beyond the retina [Hubel & Weisel 1977]. As is often the case in primate vision, there is more processing power in the system than our models yet require. But computational models are playing a stronger role in slip than in any other aspect of vision (except perhaps stereopsis) and these will be discussed below.

It is obvious from the Nyquist sampling theorem that a sufficiently rapid sequence of frames will be indistinguishable from continuous change for any visual system. The sequencing rate needed for this varies somewhat by task, but is in the range of a few milliseconds. McKee [1981], Burr, Ross and Morrone [1986] and Newsome *et al.* [1986] have interesting results along these lines. Therefore the difference between the slip and match system can not be attributed simply to continuous versus sampled imagery; finely enough sampled sequences are continuous to the system. But there are a number of characteristic differences. The long-range or match mechanism is dichoptic and can function with frames separated by many degrees and by hundreds of milliseconds in time. As we will see in the next section, there are also other reasons to believe that it is at a higher conceptual level than slip. As always in nature, it is futile to look for an absolute cut-off in the operating ranges of the two mechanisms [Mather & Anstis 1986]. They overlap significantly and usually yield consistent input to the relaxation of Figure 3.

There has been a great deal of work on the slip mechanism in psychophysics, in physiology and in computational modelling. We will begin the discussion with the models, because they provide a framework for integrating the experimental data. There are two basic kinds of slip model in the literature: delayed comparators and gradient models. There is a great deal of current work on both models and several attempts to combine the ideas. Hildreth and Koch [1987] have written a good introductory survey.

One way a system could detect local change (originally suggested by Reichardt [1961]) is to embody networks that compare nearby points of the image at slightly different times. There are considerable data [Hildreth & Koch 1987] indicating that invertebrates use this mechanism and that something similar operates in the retina of rabbits [Barlow & Levick 1965]. Motivated by experimental findings (e.g. anti-inhibition agents block directional selectivity), the standard model is a complementary pair of detectors, each of which is a veto circuit. If an image property is detected at (x,t) and not at $(x + \Delta x, t + \Delta t)$, then there is evidence that the image did not move in the $+x$ direction with speed $\Delta x/\Delta t$. The difference of two opposing comparators signals leftward, rightward or null movement. One long-standing problem for this model has been where the Δt (of tens of milliseconds) might be realized in the mammalian visual system. Mastronarde [ref] has recently found "lagged response" cells in the LGN that have properties consistent with their being the substrate of the required delay. Even without detailed calculations one can see

that any network of this sort would need unrealisitc numbers of neurons to have good resolution in space and time over a significant portion of the image. As we will see, there are a number of other reasons for believing that the short-range system provides only coarse estimates of change.

The other basic proposal for local slip detection arises from viewing the system as a truly continuous one. A continuous system could calculate local values of $\partial I(x)/\partial t$, the change in the image with time at a fixed point. We can also assume that such a system could also compute a spatial derivative $\partial I/\partial x$ or its analog in two dimensions, the gradient. If, by magic, the visual system could compute these two derivatives, then an estimate of short-range image motion (slip) is given by:

$$slip(x) \sim \frac{\partial x}{\partial t} \sim \frac{-\partial I/\partial t}{\partial I/\partial x} .$$

There is also a great deal of work on this theme and some recent efforts to build analog VLSI chips that compute continuous derivatives [Hutchison *et al.* 1988]. Notice that the equation above is poorly conditioned for image regions of constant intensity ($\delta I/\delta x \sim dI/dt \sim 0$) so any scheme should emphasize discontinuities [Marr & Ullman 1981]. It is also true that delayed comparators will give no net output for constant regions; the detectors in both directions will cancel. In fact, there are several ways of making the two styles of slip detection look quite similar [Adelson & Bergen 1985].

From a computational perspective, either style of slip detector can yield only crude noisy information about changing scenes. One reason for this is the so-called aperature problem (cf. Figure 7). A local computation of slip can yield reliable information only on one component of the underlying motion, so measurements must be combined to yield even a crude motion analysis. A great deal of work has been done on combining local slip measurements in computer modelling and some in psychophysics and neurophysiology [Adelson & Movshon 1986]. For example, there is good evidence that many cells in area MT (cf. Figure 6) but not primary visual cortex (V1) respond to the combined motion of two orthogonal moving patterns. Much of the work in computer vision has been concerned with combining many local measurements to produce an overall "optical flow" field. The basic idea is to assume that flow changes slowly along curves or within regions and to solve for best fit. From our perspective this is another local smoothing operation and it is not important that no one knows how to do it very well [Neumann 1986]. A related, and more important, computation involves discontinuities in optic flow. These can often be found with simple local detectors and provide powerful cues for segmentation. There is evidence [Allman *et al.* 1985] that many cells in MT and some elsewhere are particularly good at this. The current model assumes that flow discontinuities are important segmentation cues (cf. Section 4) but the model has no technical contribution to suggest. Similarly, this paper is not concerned with the well-studied use of optical-flow fields for general navigation and collision avoidance [Lee 1980; Lawton *et al.* 1987]. For our concerns with recognition, the critical point is that

recognition from flow fields (let alone raw motion images) has not worked very well and there are basic computational reasons [Verri & Poggio 1987] for not expecting much success along these lines.

Figure 7 about here

Although short-range retinal motion, or slip, is locally neither very accurate nor very useful for indexing, it plays several crucial roles in the overall change model. As mentioned above, local slip discontinuities are important for segmentation. Slip provides feedback to the pursuit system for zeroing relative motion of the object being tracked. Overall patterns of optic flow can trigger attention and serve as aides to navigation and other spatial tasks. And, most importantly for us, the slip information provides one with the basic inputs for change processing (Figure 3). This is discussed in detail in the next section; we conclude this section with a proposed reconciliation of continuous and Reichardt-style slip detection models.

Although the Reichardt-type and continuous models of slip detection are similar, there are a number of ways in which they make different predictions and behavioral experiments to discriminate between the two models have been attempted. For example, Mather [1984] presents evidence favoring the continuous derivative model. On the other side, Van Doorn *et al.* [1984] describe a number of experiments that they interpret as strongly suggesting Reichardt-type models. This kind of mixed behavioral evidence, along with the anatomy and physiology of primate visual cortex [Pettigrew *et al.* 1986; Maunsell 1987] suggests the following possible mixed model for slip computation.

We have already seen that there are two visual channels (P, or sustained, and M, or transient) that remain segregated through much of the visual system but also have many interactions [Maunsell 1987]. It could well be that the computation of local slip in primary visual cortex (VI) is one of these convergences. Example 1 illustrates that any system that compares distinct events confronts the matching or correspondence problem. Earlier in this section we showed how a purely local detector is inherently inaccurate, but unambiguous. We also know from apparent motion studies [Madden 1988b] that local slip is an effective cue in resolving correspondence problems. Suppose something similar was operative at a finer spatial grain in short-range motion detection. Input from change sensitive M (magnocellular) LGN cells could establish possible places and directions of local motion. This would greatly help a Reichardt-type mechanism in VI solve local correspondence problems. This model seems to be testable with currently available selective blocking techniques. Even if this particular suggestion is shown to be worthless, there appears to be considerable value in exploring interactions of static and dynamic mechanisms. The next section does this at a higher level, the perception of overall motion and change.

IV. Perceived Change and Apparent Motion

There are a great number of detailed computational models of motion and change processing, but none that attempt to integrate the varied sources of input and different causal attributions, as depicted in Figure 3. The problem is one of complexity -- it is too early to expect a detailed integrated model of change processing. But the problem with restricted models is that they run the risk of omitting crucial considerations. In this section, I will out line a computational model of intermediate complexity that attempts to encompass all the issues.

The first, and most critical, point of the exercise is to observe that the multi-input, multi-output problem posed by Figure 3 is inseparable. The change processing network (in animals or connectionist models) must reconcile all its inputs with a coherent set of hypotheses. This is exactly analogous to the processes of static vision where, e.g. the brightness at a non-specular point is a function of illumination, reflectance and incidence angle following the equation:

$$B = I \cdot R \cdot \cos \theta \ .$$

The static vision network must somehow de-convolve the right hand side to get estimates for the illumination, reflectance and local shape at each point. There is now a vast literature on how to solve these under-constrained or "ill-posed" [Poggio] problems under various assumptions. The idea that such systems can be (and probably are in the brain) solved by relaxation of a neural-net style representation goes back to [Barrow & Tanenbaum 1975]. A great deal of current work in static vision is directed at solving these multiple interacting constraint equations.

Change resolution has a similar computational character and is most likely solved by a similar relaxation procss. One should think of the visual system (and the change subsystem) as constantly trying to compute the most plausible explanation of the spatio-temporal flux and proprioceptive feedback it is confronting. As discussed in Section I, it is these constructed, distal representations that form the substrate for our actions and phenomenal perception. We see "apparent motion" when that is the best fit to a visual happening. We find it less compelling than real motion because we are unaccustomed to movement in which the slip inputs are missing. We see it differently depending on fixation, suggestion and a variety of other factors. The term "apparent motion" has become ill-defined, often being used to designate any stroboscopic change however fast. It follows from the Nyquist sampling theorem that sufficiently close frames will be indistinguishable from continuous motion and experiments have confirmed this [Burr *et al.* 1986]. It is also sometimes confounded with "induced motion" which properly refers to perception of motion of one figure due to changes in the rest of the scene. We will use the term apparent motion to denote an interpretation of a sequence of visual scenes as containing moving objects in the absence of slip cues. This is an unusual, but not impossible, situation in nature. Normally the various cues work together to produce a strong, veridical perception of what is happening. The basic operation underlying apparent motion is the

matching of two objects that appear in successive frames. There is a large literature in this area and enough data to provide a serious challenge for any model; in fact, none has been proposed. But a central aspect of our story is that the match mechanism revealed by apparent motion experiments is a fundamental component of normal change perception and must be accounted for. The second part of this section will outline a proposed model of the match mechanism and its relation to other change cues. The design and implementation of a closely related model is being carried out by Tom Olson as part of his dissertation [Olson 1988].

Another way of viewing apparent motion and the match mechanism is as one part of our general ability to interpolate and extrapolate events. We are quite good at interpolating causal chains between events that are exceedingly distant in space and time. When an apparent motion experiment uses too long a time separation, the perception is of succession without movement. An intermediate state, called ϕ in the literature, is reported subjectively to entail a definite notion of motion, but no particular trajectory. There is also a situation in which we get apparent movement from one image. Given a static image with motion cues (such as a person running), people will often confuse this image with one that would be later (but not earlier) in the sequence [Freyd, 1983; Finke *et al.* 1986]. For these reasons, among others, we postulate that there are three change mechanisms with overlapping ranges in space and time. Figure 8 is an informal idea of how all this is supposed to work ([Madden 1988b]).

Figure 8 about here

One basic question about matching concerns whether local or more global features are used to match two objects. An intuitive solution can be derived from considering the apparent motion sequence:

Example 3:

All of the local features are the same for both objects, but we always see movement from 1) to 2) as including a rotation -- the overall shape is critical in matching. Experimental confirmation for this intuition can be found: e.g., [Ramachandran *et al.* 1973] who found good apparent motion between squares defined by texture differences in images with no shared texture elements. From this perspective, a wide variety of experimental results fit the same general pattern. Lin Chen [1985] has shown a general preference for topological matching and Watson [1986] suggests

other criteria. For example, coarse patterns are known to dominate finer ones in motion perception [Ramachandran & Cavanaugh 1987]. Also apparent motion between objects defined by subjective contours or local-motion differences [Ramachandran 1985a] fits the idea of matching large-ish "objects" at a fairly high level in the visual system. This also seems to be a good way to look at the "vector analysis" proposed by Johannsen [1973] to account for our ability to recognize people from just of the motion of lights at their joints. The overall body motion is recognized and other motions are seen as relative to the main one. Notice, by the way, that the need for relative motion calculations preclude extending the pure signal processing models of motion analysis beyond the slip domain of Section III. Relative motion responses of various kinds have been found in appropriate areas of visual cortex [Adelson & Movshon 1986; Allman *et al.* 1985].

Another preliminary question concerns the ecological role of the match mechanism. If slip is almost always present, why do we need matching? The most commonly stated use of the match system is for occlusion; it is clearly useful to be able to predict the trajectory of something passing behind a visual obstacle. The general notion of object constancy through matching is important in several other tasks, including stereopsis and reorienting after saccades and head or body movements. I suggest that the match system operates continuously and plays a critical role in normal motion and change perception [Prazdny 1986]. For one thing, the local slip mechanism is not good at discriminating trajectories, especially at high speed. There isn't enough information available locally to make a sharp distinction about the direction or speed of motion. Nor will summating local measurements suffice, particularly in complex scenes with multiple motions. The claim is that an inherent part of change processing is to match moving objects at (Δx, Δt) values large enough to allow for accurate trajectory calculations. A fairly detailed proposal of how the match system might work is contained in the second half of this section. What is important here is that the match mechanism works on "segments" however defined and interacts strongly with the other major cues for motion and change including slip (Section 3), context and eye-movements. All of this can be seen as part of the giant relaxation postulated for the stable feature frame in Figure 1. In Figure 9 we show how the change sub-system of Figure 3 fits into the four-frames model. The motion and change network of Figure 3 is seen to be an elaborated version of the "motion" stable feature frame. At one level of analysis, computing perceived motion is of the same character as computing perceived color, shape, etc. The two-way indexing-context connection to world knowledge is essentially the same as in Figure 1, but is depicted a little differently.

But, there are also important differences between perceived change and other constancies. One complication is that while object motion is a well-defined visual primitive, the change network also detects alteration in all the other visual features such as color, depth, etc. Section V of this paper suggests how the other constancies of the stable feature frame can accommodate motion and temporal change. Another important uniqueness of the change system is the role of eye-movements. Much of the work in Feldman [1985] was concerned with fixations and how stable feature

maps might be maintained. There was no serious mention of pursuit eye movements, but they are essential to the current story.

Obviously enough, we should not attribute external motion to retinal slip caused by our own eye or body movements. There are known to be at least two major kinds of cues involved in distinguishing self from object motion. One source of information is extra-visual, including vestibular and efference-copy cues, sound, the feel of wind, etc. These could be included as context in Figures 3 and 9 or could be added as another class of inputs. Also included here should be the distinction between purposive pursuit eye movements and compensatory reflexive smoother movements [Post & Leibowitz 1985]. The other source of self-object motion discriminations is in the optical signal itself. A large uniform flow field in the periphery is a strong cue for self-motion, as everyone has discovered from moving trains, etc.

Much of the rest of the paper is concerned with fleshing out the implications of Figure 9. The next part of this section focuses on suggested motion primitives (for indexing into world knowledge) and how they might be computed. Section V looks backward to the other stable feature frames and considers how they might deal with motion and change. In Section VI, we return to indexing and world knowledge -- with the representation of dynamic models a central issue. Finally, Section VII considers self-motion and the exocentric environmental frame. Some notion of how the "what" and "where" systems might interact conclude the marathon.

A Connectionist Model for Perceived Motion

The first step in building any model is to specify its inputs and outputs. In this case, the output representation is critical because it is the link between visual change detection and the recognition of occurrences in the world. Recall that the entire change perception mechanism is part of the Stable Feature Frame and that its output will be used as part of the indexing process that accesses world knowledge (cf. Figure 9).

The output of the perceived change network will be in terms of *trajectories*. A trajectory, more precisely a trajectory segment, will be a pair consisting of a circular arc (with a straight line as a special case) and a constant speed.

$$t = (a,s).$$

This particular representation was chosen because it is simple and tractable and because a wide range of data suggests that primates deal with motion in a way that is consistent with this form. For example, people [Todd 1981; Lee *et al.* 1983; Ramachandran & Anstis 1983] and animals [Lee & Redish 1981] appear to normally use a constant speed assumption in extrapolating motions. Psychophysical evidence [McKee & Welch 1988] suggests that the speed is in angular not constancy-corrected metric units. Notice that constant speed along a curved trajectory is not constant velocity. Our respresentation is such that any change in speed or any deviation from

the current path requires a new trajectory segment. These second order changes turn out to be important indexing cues as we will see in Section VI. There are also undoubtedly higher order motion primitives such as rotation and expansion patterns used in indexing, but these are not considered here.

The output space of trajectories must be explicitly represented, as always in connectionist models. We will suppress efficiency questions for now and assume (counter-factually) that there are enough units to represent every trajectory of interest. For concreteness, let us suppose that each circular trajectory segment is represented by its center point, radius, start point, ± arc-distance and speed. In two-dimensional space this would require six parameters if we assume the start point is coded as a position on the circle, or five if it is represented directly. Straight line trajectories could be represented similarly as a line, start point, distance and speed. Of course, one really needs three-dimensional trajectories and the parameterization gets out of hand quickly. This is a standard problem in connectionist models and Olson {Olson 1988} has some specific suggestions for efficient encodings for this domain. Any such encoding will entail an inability to make certain simultaneous judgments and thus model (more or less successfully) various experimental findings.

A crucial assumption shared by Olson's model and mine is that the match system can only focus on a single parameterization of motion at one time. The results of Ramachandran [1985b; 1986] suggest that a single global representation for the parameter set: (speed, radius, ± arc distance) is part of the representation. Ramachandran found that ambiguous apparent motion displays like the arrays of the following

Example 4: ① ②

② ①

would be perceived as all horiziontal or all vertical motion and could all be switched by changing one subarray. Other experiments suggest additional constraints of the trajectory representation, but the details are beyond the scope of this paper (cf. {Olson 1988}).

The important points here are that constant-speed trajectory segments are the output of change percception and that they are represented explicitly. This allows us to employ the standard connectionist device of having the various trajectory-units compete to explain the different pieces of evidence presented as input to the match subnetwork. Again, the details require at least a thesis worth of work but the general outline can be seen already. Consider the simplest kind of apparent motion experiment, with two identical dots presented in the sequence:

Example 5: ① ②.

The obvious interpretation of horizontal movement to the right at a speed $s = \Delta x/\Delta t$ is the dominant one for a wide range of Δx, Δt values, as expected. But under various conditions {Madden 1988a}, people see other trajectories, usually roughly circular arcs either in the plane of the display or orthogonal to it. No one ever sees more than one path, strongly suggesting mutual inhibition.

One condition that leads to a curved path percept is a $\Delta x/\Delta t$ ratio that would require the straight line speed to be too slow (for this observer's network) but is appropriate for a longer path. Shepard and Zare have shown [1982] that flashing a curved blur-path between ① and ② very reliably induces a curved path percept. An obstacle in the straight path can induce the curved perception. Suggestion is also quite effective as is adding a frame or two of slip cues after flashing ① {Madden 1988a}.

The computational idea that arises from all this is the standard one of competing trajectory units, each receiving activation from a variety of sources. Obviously enough, priming, blur paths and slip cues could be directly wired to all trajectory units that are consistent with that input source. An obstacle could inhibit paths that go through it. A connectionist rendition of the match process is more complex and more interesting.

First, consider the case where all of the image elements are identical dots, a case that has been extensively explored experimentally [Kolers 1983, 1972]. The first and most serious issue is that any match process would seem to call for a buffer of the information in the first (or nth) frame while the next frame was being processed. This is not, of course, restricted to apparent motion -- any feature matching implies relatively low-level memory and there is no physiological evidence for multiple buffer-like storage in the visual system. Olson {1988} has developed an ingenious scheme for matching dot images, based on the quantitative decay of activation at image points and comparator circuits sensitive to time and distance separations. Some such mechanism will be needed and none of the slip detection techniques of Section III will suffice -- matching can take place with separation of many degrees and hundreds of milliseconds. For now we just assume that some matching scheme sensitive to time and distance is realized and that its results are also fed to the appropriate units of the trajectory network.

One controversial issue is whether the proximity component of the match competition is based on 2-D or 3-D distances. Different experimental paradigms have yielded opposing results. There is a simple reconciliation of most of these results in our current models. When 3-D information is reliably present (e.g. [Green & Odom 1986]) it is used in matching. Attempts to induce three-dimensionality by using figural cures (such as a coordinate frame) do not work well because the match calculation is at a lower level and apparent motion conditions do not provide enough time for context to be established. A major goal of Olson's work {1988} is to fit as many experimental findings as possible.

But matching is not restricted to just dots and our story must be elaborated. Consider the earlier example of the sequence of rectangles.

With appropriate timing, this is always seen as the continuous rotation of a rectangle while moving along a curved path. The current model treats this, and similar phenomena, as a separate but related set of parameter fitting relaxations. Our assumption is that the objects 1 and 2 are represented in the Stable Feature Frame as a vector of properties, as in the original model. If objects 1 and 2 are matched (by a complex relaxation), the discrepancy in orientation activates the unit representing say, a clockwise rotation of 90°. This unit is consistent with an upward circular arc and sends activation to it so that we "see" the combined motion. According to the literature [Kolers 1983; 1972], some discrepancies (size, orientation, affine change) yield continuous perceptions where others (color, topology) do not. As these relations are worked out, they will inform the detailed construction of future models.

The final question involves structured objects: How could they be matched? Consider first a simple case.

Example 6:

Under appropriate timing conditions, this will usually be seen as a clockwise circular motion with a 180° rotation. This suggests some elaborate shape comparisons in the match system and led us to look further {Madden 1988a}. It turns out that the match is simpler than that and will often produce physically unlikely perceptions. For example, the sequence:

Example 7:

can be seen as a rotation towards the viewer, but often is not. With the appropriate timing, one sees a large rectangle sliding to the right with the small tab moving (in a

vague way) to its new position. This and a number of other results [Ramachandran *et al.* 1983; Mather & Anstis 1986] suggest that the match system works primarily on large objects (areas with significant low spatial-frequency energy). There is no reported case where the large object does not get matched coherently. The general notion that the change processing system favors explanations with large coherent object motions is ecologically plausible and is consistent with many experiments. I do not know of any experiments suggesting otherwise.

We are now in a position to consider change in more complex situations like a trotting horse or a crowd scene. As we will see in Section VI, recognition of complex motion is presumed to be heavily dependent on having available models to integrate the individual trajectories. Subjects who easily perceive moving light display presentations of moving people will totally fail if the displays are inverted [Sumi 1984]. The match system being modelled here can only compute trajectories for a few points of interest at a time. It does need the ability to compute trajectories relative to frames that are themselves moving and mechanisms for this are being worked out by Nigel Goddard as part of his forthcoming thesis (cf. section VI) [Börjesson & Van Hosten 1975].

V. Temporal change and the stable feature frame

The cornerstone of the four frames model was a postulated mechanism, the stable feature frame, that both computed perceptual constancies and accumulated visual information across saccades. This depended in a critical way on a static world -- the observer's eyes were the only moving thing taken into account. Information from different saccades was essentially overlaid on the stable feature frame like a collage. The mechanisms suggested there simply do not work for a moving observer or objects, as was pointed out. The proposed spatial layout of the stable feature frame was also the most controversial aspect of the entire model and it is worth a brief reprise of the issues and how they might affect our current concerns.

The fundamental problem being confronted is how we maintain the perception of a stable, uniformly sampled visual world despite the exponential fall-off of resolution and some four saccades per second. Information is obviously integrated over saccades and there seem to be two basic possibilities, classically known as pre- and post-categorical integration. The post-categorical integration hypothesis is that each fixation causes separate visual processing up through the levels where objects are labelled or categorized. In the four frames model, this would be the world knowledge formulary (WKF). Successive saccades are integrated at this symbolic level. Top down feedback from concepts to early vision would be retinotopic. The main argument for this view is that all visual neurons show some retinotopic effects. The earlier visual areas are clearly retinotopic, but there are plausible treatments of higher-order visual areas that are consistent with a head- or world-centered coordinate frame [Anderson 1987]. The post-categorical integration hypothesis has never seemed plausible to me, but I have been unable to devise or elicit definitive experiments. There is currently some interest in behavioral experiments that bear on this question [Irwin 1988], but they are mainly testing the hypothesis that raw visual information is being combined across fixations. Neurophysiologically, the spatial integration of features across saccades should be subserved by the more posterior parts of infero-temporal cortex, which are just beginning to be explored.

One version of the alternative, pre-categorical integration notion was defended in [Feldman 1985]. Integration was claimed to take place in a roughly head-based spatial frame and to be at the level of distal (constancy, intrinsic) features. The details of this proposal can be seen in Figure 10 and are quite straightforward. The mapping from retinal to head coordinates is mediated by the current value of gaze. Computationally, this is a simple 1-many map where at each time only one destination is enabled. This is a reafferance proposal [Haber & Hershenson, p. 209] but is the only one I know of that makes explicit how the anticipated and received information is to be combined. In my model, the stable feature frames are always integrating and normalizing lower level features and this (among other things) provides the substrate for visual stability. Figure 10 also suggests how the fish-eye magnification factor of the retinotopic frame could be mapped to more uniform later representations; [Feldman 1985] also suggests further properties of this construction

for integrating saccades. Reciprocally, matching expected features helps compensate for inaccuracy in saccades.

Figure 9 about here

Introducing a moving observer and objects complicates this story in several ways. Wallach [1987] reviews the evidence that people are also very good at anticipating scene changes based on their own normal head or body movements. But if you view the scene behind you (with a mirror) while walking forward, things appear to shrink or recede. There seems to be no inherent difficulty in extending Figure 9 to deal with such findings, but it would involve inputs from the world-knowledge and environmental frames. The anticipated appearance features of objects as one moved to new positions would still be set in the stable feature frame. Objects that themselves move (and interact) present a more difficult set of problems.

Figure 10 about here

As we have seen, the visual system incorporates elaborate mechanisms for tracking an object of interest and keeping it stable on the retina. When an object is being tracked, our ability to recognize other things in the scene is greatly reduced. An easy demonstration of this can be achieved by tracking your finger across a piece of this text. You can also choose to read the text using your finger as a moving pointer. This suggests part of the answer; one important way to recognize moving objects is to track them so that all the static recognition mechanisms can continue to function. Notice that in this case retinal slip does not indicate object motion but pursuit error on the object being tracked. Slip of the opposite sign and equal magnitude to the tracking trajectory is a strong cue for non-motion in the background. This context-dependent interpretation of slip information has been discussed in connection with Figures 3 and 9.

Under post-categorical integration, the processing of a tracked object would seem identical to a static one, except for slip cues. It is not clear how one would account for the perception of a stable background, which you can notice as you track your finger.

The stable feature frame model has no trouble keeping track of the stable background (at coarse resolution). The gaze mapping that relates eye to head position would simply work continuously instead of discretely as it does for saccades. But the exact treatment of the object being tracked becomes somewhat problematical. The direct realization of tracking would have the features of the tracked object over-write those of background objects in the target passed over their positions. The spatially-independent feature conjunctions nodes (cf. Figure 2,

bottom row) would continue to get the same activation patterns, because the same feature co-occurrences would hold, albeit at different positions. The model already included the notion of extra activation for the positions being foveated and this would also presumably move smoothly with gaze. There is a potential problem in that normally the stable feature frame is integrating the features of spatially adjacent points and this won't work with the tracked object. It could be that depth differences inhibit these neighborhood processes as they appear to do in other cases, particularly occlusion. The alternative model is that the processing of a tracked image remains separate from the background. An experimental test of the alternatives would be the extent to which a tracked object can be followed through camouflage of approximately the same depth.

But the case of a perfectly tracked rigid object is the simplest temporal issue for the stable feature frame. There is, in general, no way to prevent retinal slip from being an important visual parameter. Consider the problem of viewing a galloping horse, perhaps to see if the jockey is performing well. One can, and does, track the horse's body, but there are still all the relative motions of parts and these cause continuous change of the retinal image. This causes several problems for any model {Goddard 1988}. One of the easier problems is maintaining the position of the horse in allocentric space; this can be computed from its depth along with head and eye tracking data and can be updated in the environmental frame. All this will be discussed further in Section VII.

The basic temporal issue in feature calculations is how the relaxation networks cope with constantly changing inputs. This is the same problem for the pre- or post-categorical integration models. If the shape, say, of a sub-object is computed by neighborhood interactions, what happens if the input changes faster than the relaxation time? The obvious technical solution of temporal discretization (i.e. a shutter) is ruled out by what we know about the early stages of vision (cf. Sections II & III). One mechanism that could work is for each sub-object of a complex object to have a "resting position" with respect to the larger object. For example, the resting position of a horse's legs with respect to its body could be straight down. All of the static properties of the part: size, shape, color, could be computed only in the resting position and the relative motion trajectories computed separately. The easiest way to do this would be to gate, e.g., color, to the appropriate feature units only when the sub-part was in its rest position. Of course, all of this depends on having segmented a complex moving object into appropriate sub-objects. As before, the model assumes segmentation itself is part of the giant relaxation process and uses motion as well as static cues. As noted in Section III, discontinuities in the slip field provide a powerful segmentation cue.

The original stable feature frame also allows for different points of fixation while tracking a complex moving object. The roughly head-based spatial layout can retain feature vectors as in the static case. Changes in the point of fixation from imperfectly tracking, compensatory saccades or information demands can be mapped

appropriately. Some such mechanism is required to account for people's ability to detect tracking errors in the absence of slip discrepancies [Collewijn 1986].

It should now be clear why the change processing sub-network of Figure 3 and Section IV is viewed as part of the stable feature frame. Motion features, in distal or constancy form, interact with the static structural invariants in a huge deconvolution network. From a neuro-anatomical perspective, these interactions would be mediated by e.g., the bi-directional links between areas V4 and MT. Other relations between the putative "what" and "where" systems will be discussed in Section VII. The trajectory information which is the presumed output from the change subsystem is a key input for recognition in the world knowledge formulary, as described in the next section.

VI. Representing and using change

The paper so far has said very little about how the change parameters, extracted in intermediate vision, are put to use. The major uses discussed were feedback to control systems like pursuit and to other intermediate calculations like segmentation and depth. In this section we focus on the most important use of change information -- recognition. As in the earlier four frames paper, the central problem of vision is taken to be *indexing* -- utilizing the results of intermediate vision to recognize (or categorize) the visual world. A significant extension to the previous treatment is a discussion of recognizing what objects are doing in addition to what they are.

Of course, recognizing characteristic motions is one way of categorizing certain complex objects like animals or machines. The most striking example is recognition from pure motion cues in moving light images [Johannsen 1973; 1975]. Given a sequence of images containing only the motion of lights on a few joints, subjects are able to quite rapidly identify many human motions and even determine individuals. There are several computational models of how trajectories might be calculated from moving light sequences [Rashid 1980; Hoffman & Flinchbaugh 1982], but none that suggest how recognition is achieved. A first step is to postulate a representation of world knoweldge that incorporates the information needed for indexing by motion cues.

Figure 11 is a connectionist model of a horse, essentially the same as that used in Figure 2 and similar to hierarchical descriptions commonplace in computational knowledge representation systems. The horse shape, in a particular view, is subdivided into sub-parts that are further subdivided, etc. What is new here are the diamond-shaped nodes linking subparts such as head and neck. The representation is based on the notion that for articulated objects like the horse, there is a specific point at which subparts can be considered to attach. For each of these, the diamond node describes the permissible variation in angles for this connection. The angles are represented relative to an axis perpendicular to the major axis of the base part [Hrechanyk & Ballard 1982].

Figure 11 about here

There are a number of technical details to be explicated, but the general idea should be clear. Our representation of articulated objects now includes the allowable range of attachment angles. This can be used for recognition either in verification or, if angles can be computed in the stable feature frame, as part of the initial indexing. Notice that the principal-views representation allows all of this to be two-dimensional. One could also add a stiffness parameter that defines the allowable

bending of a component's major axis. This captures much of the kinematics of articulated objects, but we must also deal with dynamics.

In order to describe the motions of a complex object (or group of objects), the representation must be extended to explicitly include temporal change. We will use the term "scenario" to refer to both the motions of a single object and to groupings. A typical scenario might be a gait (trot, canter, etc.) of the prototypical horse on your walk to lunch. An idealized gait can be specified in terms of the relative timing of joint motions. As in Section IV, we assume that all motions are represented as traversing a curvilinear (here circular) path at a constant speed. Since each joint has a fixed path, one needs to specify only the speed of motion and relative timing (phase). If we also specify when and where reversals of directions take place, the representation is complete.

The complete specification of gaits and other scenarios presents significant technical problems in connectionist modelling. There has been connectionist work on sequences and loops [Addanki 1984; Chun 1986; Kleinfeld 1987] and a great deal of other neural modelling of animal motor control, but the needs of a recognition system are different. For example, we can recognize motion (of a horse or machine) that we have no way to carry out ourselves.

Nigel Goddard is developing a connectionist motion recognizer as part of his doctoral dissertation. The following example is taken from his work. Figure 12 depicts a periodic motion of a 4-stick articulated object as shown vertically in the right-hand column. The first three columns show how each of the three "legs" move with respect to the horizontal "body" which does not move in this example. Sticks 1 and 2 rotate 360° clockwise out of phase and stick 3 oscillates over a 180° range. The idea, of course, is to use the local motion sequences in the first three columns to index (recognize) the complex motion of the entire object in column four.

A central idea in Goddard's model is to rely heavily on points of disruption of simple trajectories -- these correspond to points of reversal or acceleration. In Figure 12, the key points of motion change are shown in heavy lines. Rods 1 and 2 have key points at steps 0, 2, 5 and 7 and rod 3 at 0, 3,5, 8 (step 10 is just 0 again). The internal representation of the motion sequence of Figure 12 is given in Figure 13.

Figure 12 about here

Row a of Figure 13 shows the three joints and the directions in which angles are measured. The time sequence of key points for each joint are given in rows b and c of Figure 13, with each key point labelled by the step at which it occurs for clarity. Row c shows the simultaneity constraints which actually encode as two-way links the rule that this motion requires the coordination of the separate joint motions. Row b of Figure 13 depicts the unidirectional links that capture the relative order

constraints among the motions, both for a single joint and between different joints. Thus the representation is independent of overall time scale and can be matched by any motion with the right relative timing. Characteristic motion sequences like Figure 13 are added to the static and kinematic characteristics of Figure 12 to complete the description of an articulated object.

Figure 13 about here

Figure 14 presents some technical details on how motion matching actually takes place in Goddard's system. The general approach follows our techniques for static matching [Cooper & Hollbach 1987]. Basically, the network needs to bind each individual piece of input information to an appropriate piece of the model in such a way that critical relationships are preserved. When, as usual, there are several possible models to match, the one with the best overall fit to the input data is selected. In Goddard's example each input is characterized by one static feature, joint angle, and one dynamic feature, ± speed, along its circular trajectory. This is the kind of input one might expect as output of the networks discussed in the previous sections.

The upper box of Figure 14 depicts the instantaneous input of one joint, an angle of 270° moving at +22.5°/second. The unit E detects an "event," a change in the active speed unit from the row above. The leftmost network of Figure 14 is just a copy of our model of joint 1 (the left hand column of Figure 13). The lower right network is the "binder" which encodes the compatability between the input (above) and the particular model component to its left. The unit labelled R_1 is a sub-binder explicitly linked to the feature (270°, 22.5°/second). If these are present in the input at a given time, R_1 will be activated and this will send some activation to the model unit labelled 0. If the sequencing is correct, unit 0 will send activation back to the central binder unit B and this will, in turn, enhance the activation of sub-binder R_2 if its input configuration happens to appear next. The positive feedback between expectations from the model and compatible data will lead to high activation for the appropriate bindings. The network can produce good matches between individual model and image motion features, but additional mechanisms are necessary to capture phase relations among joints. Even in our tiny example, joints 1 and 2 are the same except for phase with respect to joint 3. Goddard's model is able to handle problems like this and is being extended to more complex matches.

Figure 14 about here

The point of all this is to suggest that change information can be incorporated as another set of recognition cues. The static recognition network proposed in [Feldman 1985] and implemented in [Cooper & Hollbach 1987] has been extended by

Goddard without major change to incorporate dynamic data. Even the current early state of development seems adequate for representing and recognizing well-known motions like Johannsen's examples or a horse's canter.

All of the discussions so far have concerned the motions of more less complex single objects. The model also treats the problem of representing and using change information about larger configurations. The static version of this problem was handled in [Feldman 1985] by postulating "situation" networks that represented collections of related objects in a situation like someone's office or Harvard Square. Situations have the same formal structure as complex objects and were assumed to be recognized in similar ways. The representation was assumed to have several levels of detail and to interact with the environmental (allocentric) frame in ways that will be reviewed in the next section.

Essentially all of these ideas carry over to the dynamic case. The static notion of situation is extended to include change and is renamed as the "scenario." A scenario is a situation graph with associated change constraints as in Figure 11 and a trajectory network like that of Figure 13, but it obviously does not need to be periodic. Scenarios for your office might include some allowable displacements of objects and trajectories for one or more common motion sequences such as answering the phone. Scenarios can nest as situations do -- for example, the detailed motions required to pick up the phone. The current model does not cover extended time sequences such as plans or stories, but these would presumably be constructed from nested scenarios, among other things. Recognition of scenarios involves both static and dynamic cues, exactly as in the detailed example above. Once activated, scenarios are assumed to have the obvious context effects.

As with static situations, scenarios are the basic mechanism for establishing coherence in the model. Both visual data (frames 1, 2) and one's position in space (frame 4) can change rapidly, but a known scenario remains current for an exended time. Scenarios can be used to incorporate information about different objects and their activities. Most important, they can be used to predict and respond to external events. All of this is done in connection with the environmental frame, which encodes relations in external real-world coordinates and is discussed in the next section.

VII. What and Where

Virtually none of the discussion thus far has dealt with the realities of external, three-dimensional space. The low and intermediate level mechanisms of the early sections, the recognition networks and even the scenarios described in the previous sections are all in the mind of the observer. There obviously must be a connection to the external world and this modeled (both in Feldman [1985] and here) as the fourth or environmental frame.

The environmental frame is assumed to be a farily direct encoding of the space surrounding the observer. There is by now a wide range of converging eficence (e.g., [Kosslyn *et al.* 1988], Figure 6) that the visual system has two distinct (but interacting) pathways often denoted by "what" and "where." The "what" or ventral pathway is specialized for form and other properties that are central to recognition; most of Sections IV-VI were focussed on this pathway and constitute a model of its functioning. As was mentioned in the Introduction, it appears to be computationally necessary that the complex networks needed for recognition be independent of spatial postion – otherwise far too many units and connections would be required. It is widely believed that the dorsal, "where," system complements the ventral system by processing location information. Kosslyn *et al.* [1988] present a block-diagram level model of the two systems based primarily on studies in neuro- and cognitive psychology. Our concern here is at a lower level of detail -- what computational mechanisms could support the functions of the two sybsystems and their interactions. It turns out that the resulting structure differs somewhat from that of Kosslyn *et al.*, but in relatively minor ways. I will first describe my computational model of the external environment and then relate it to a number of issues concerning vision and space.

The environmental frame in the model is intended to correspond roughly to posterior parietal cortex. It is in allocentric (exocentric) coordinates and has variable resolution representations for different scales. The model uses rectilinear coordinates for simplicity, although this is unnatural (see Figure 15). The crucial role of the environmental frame is to ground world knowledge in physical space. In the original four-frames model, exactly one situation network in the world knowledge formulary was mapped to the environmental frame at a given instant. Some consideration was given to discrete changes in this mapping, but none to continuous change. This section attempts to address such concerns and how they interact with "scenarios" which are the temporal extension of situations.

For technical reasons to be discussed, we allow only a one-way mapping from the environmental frame (EF) to a given situation (or scenario). Links from a given position-unit in the EF would go indirectly to every object that could be in that position in some scenario (cf. Figure 15). All such links were gated through conjuctive connections where the enabling signal came from a fixed unit for each situation (now scenario). For example, Figure 15 depicts a situation where a "Harvard Square" situation is mapped to the EF. Focusing attention on some

direction and distance in the EF, combined with the activity of the <situation #> unit activates the Coop door unit in the current situation. The reason that the links in the other direction were omitted is that having them would lead to a huge number of inputs to EF units and possibly make them sensitive to low levels of widespread noise. Large fan-out does not create the same problems. The restriction to one-way situation links in the model has testable behavioral consequences, but these have not been tested. The model suggests, for example, that it should be harder to go from an object in a scenario to its absolute location than to predict an object from a location. It also suggests that attention is required for the former, but not the latter task.

Figure 15 about here

Motion and other continuous change in the real external world must also be accomodated in the environmental frame. To a first approximation, temporal change without a change in position (rotation, vibration, etc.) can be ignored in the EF and treated as changing properties detected in the stable feature frames and remembered as part of the current scenario in the world knowledge formulary. This is basically what the previous section covered. But motion in space, of the observer and other entities, is a major responsibility of the environmental frame.

Self-motion was already given a special role in the 1985 paper. The position and orientation of the observer in the current environment was treated as a special subnetwork and was updated with each motion; continuous change adds no basic considerations. In the early paper, there would be a discrete situation change at specific points in a given situation. Typical examples include turning a corner, passing through a door or switching ones level of consideration to finer detail. The idea there, which still seems viable, is that the situation (scenario) switch happens mainly in the symbolic world knowledge formulary. On a situation switch, the environmental frame must update the ego position and the mapping from environmental frame to world knowledge formulary now is gated by a different situation mode (cf. Figure 15 again). But basically no changes are needed from the 1985 version to accomodate motion of the observer.

Of course, the observer is not the only thing that moves. Perhaps the major role for spatial reasoning is predicting motions and collisions. There is considerable evidence [Lee *et al.* 1983] that some primitive flow-based collision avoidance is evolutionarily wide spread and present in humans at a very early age. This is assumed to be largely sub-cortical and not part of the four-frames model. But we also have very sophisticated abilities to understand and deal with scenarios involving moving objects and agents. The model suggests that our ability to deal with complex scenarios depends (as always) on having structures (in the world knowledge formulary) to accommodate the complexity. We will first consider how the environmental frame and world knowledge formulary could interact with motions in

a known scenario. The whole point of knowing a scenario is to be able to predict future events. Again, the bulk of the processing will be in the world knowledge formulary which has mechanisms for simulation (cf. section VI). If the situation gating-links can be kept current, activation of any position node in the environmental frame will lead to activtion of the object that should be at that position now. Predictions can not be handled this way because there is only one "now" and one current environment. All this machinery could be dedicated to simulating a scenario (and presumably is in dreams, etc), but not while dealing with the present. What seems to be needed is an additional mechanism for projecting trajectories. Recall that our treatment of motion is based on smooth, constant-speed trajectories. It is not hard to envision that given a 3-space trajectory, a speed and a starting point in the environmental frame, the extrapolator network can predict where in the environmental frame the moving object would be at various times in the future. We assume there is only one such network; there is no evidence that we can predict multiple trajectories and it would be computationally awkward to try.

This also provides an answer to how the model might deal with motions that do not fit a known scenario. The 3-space trajectory mechanism can still be used to predict the motion of a particular object. Its trajectory would be estimated in the retinal and feature frames, matched against a generic trajectory representation in world knowledge and mapped to the appropriate position in the environmental frame. Tracking an object both maintains retinal stability and helps predict the trajectory. By attending to the target object, one could predict where it is going without any prior knowledge. The advantage of having a stored scenario is that prediction can be done for several objects, including those not in view. Again this is considered continuous with our general ability to predict events that are not presently visible.

The preceding outline of environment frame function is even rougher than the other parts of the model, but the basic ideas seem plausible. An explicitly spatial frame seems necessary and sufficient for dealing with objects in the external world and their motions. Notice that the "where" system has become a "whither" network, crucially concerned with the future. This seems ecologically plausible and consistent with the notion that the dorsal pathway is concerned with change.

The remaining question for this section is how the ventral and dorsal systems might interact. The dichotomy into what and where is attractive, but does not hold up under scrutiny. For example, recognizing a complex object (horse, boss' expression) depends not only on features, but on their relative positions and distances. So "where" information is critically needed by the "what" system at an intermediate stage. Similarly, recognition from motion (discussed at length in Section V) requires trajectory information in the "what" system. In the other direction, the basic *match* process (as in apparent motion, Section IV) requires some form matching in the "where" system. For me, all of this fits into the same general story as the two kinds of slip detection in Section III. At all levels, the computational task calls for an interaction of both form-dominated, static, and dynamic kinds of

information. The physiological pathways linking the two pathways at many places from the retina onward are consistent with this story.

One basic problem for any model or theory of vision is how the what and where pathways reconverge. In the current model, they interact early and often, but there remains the problem of finally determining exactly what is exactly where in a situation. For us there are two possibilities: If one object is fixated or being tracked, the correspondence between what and where is maintained by the current state of the system. There is one most active object and one most active (perhaps changing) position. The only way that multiple correspondences between objects and positions can be maintained is by using memory structures: a scenario in the world knowledge formulary and its mapping to the environmental frame. These claims should also be testable.

Acknowledgements

The following people provided helpful comments on even earlier drafts of this paper: Chris Brown, Bob Emerson, Nigel Goddard, Mary Hayhoe, Christof Koch, Steve Kosslyn, Brian Madden, Walt Makous, John Maunsell, Suzanne McKee, Bernd Neumann, Tom Olson, Tanya Pasternak, David Plaut and Dave Williams.

Appendix A

Phenomena:	Sections
System components slower than response	(1,2,3,4)
Inseparability at the receptor	(2,3)
Slow response of receptors	(2,3)
Motion smear	(2,3,4)
X-Y, Magno-Parvo, etc.	(3,4)
Adaptation, stabilized images, small movements	(2)
Masking and meta-contrast, transient and sustained	(2,3)
Models of short range: Gradient vs. Reichardt	(3)
Cortical motion cells	(3)
Optical flow, aperatures and all that	(3,4)
Short and Long Range mechanisms	(4)
Temporal judgments	(2)
Apparent Motion	(4)
Spatial and temporal integration	(5)
Temporal change in stable features	(5)
Moving Light Displays, vector analysis	(6)
Representation and Matching of moving objects	(4,6)
Scenarios	(6,7)
Event perception	(6,7)
What-Where systems	(2,3,7)

Bibliography

Addanki, S., "A connectionist approach to motor control," Ph.D. thesis, Computer Science Dept., Univ. Rochester, 1984.

Adelson, E.H. & Bergen, J.R., "The extraction of spatio-temporal energy in human and machine vision," *Proceedings*, IEEE workshop on Motion Rep. and Analysis, Kiawah, SC., May 1986.

Adelson E.H. & Bergen, J.R., "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A, 2*, 2, 1985.

Adelson, E.H. & Movshon, J.A., "The perception of coherent motion in two-dimensional patterns," in *Motion: Representation and Perception*, N.I. Badler & J.K. Tsotsos (eds.), Elsevier Science Publishing Co., Inc., 1986.

Allman, J., Miezin, F. & McGuinness, E., "Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons," *Ann. Rev. Neurosci. 8*, 407-420, 1985.

Altmann, L., Eckhorn, R. & Singer, W., "Temporal integration in the visual system: Influence of temporal dispersion on figure-ground discrimination," *Vision Research, 26*, 12, 1949-1967, 1986.

Andersen, G.J., "Perception of self-motion: Psychophysical and computational approaches," *Psychological Bulletin, 99*, 1, 52-65, 1986.

Andersen, R.A., Essick, G.K. & Siegel, R.M., "Neurons of area 7 activated by both visual stimuli and oculomotor behavior," *Experimental Brain Research, 67*, 316-322, 1987.

Anderson, R.A., Essick, G.K. & Siegel, R.M., "The encoding of spatial location by posterior parietal neurons," *Science, 230*, 456-461, 25 October 1985.

Anderson, S.J. & Burr, D.C., "Spatial and temporal selectivity of the human motion detection system," *Vision Research, 25*, 8, 1147-1154, 1985.

Anstis, S.M., "The perception of apparent movement," *Phil. Trans. R. Soc. Lond. Series B, 290*, 153-168, 1980.

Anstis, S.M. & Mather, G., "Effects of luminance and contrast on direction of ambiguous apparent motion," *Perception, 14*, 167-180, 1985.

Anstis, S.M. & Ramachandran, V.S., "Visual inertia in apparent motion," *Vision Research, 27*, 5, 755-764, 1987.

Anstis, S.M. & Ramachandran, V.S., "Entrained path deflection in apparent motion," *Vision Research, 26*, 10, 1731-1739, 1986.

Anstis, S.M. & Ramachandran, V.S., "Kinetic occlusion by apparent movement," *Perception, 14*, 145-149, 1985.

Arbib, M.A. & Hanson, A.R. (eds.), *Vision, Brain, and Cooperative Computation*, Bradford Books/MIT Press, 1987.

Arend, L.E. & Timberlake, G.T., "What is psychophysically perfect image stabilization? Do perfectly stabilized images always disappear?" *Jour. Optical Soc. America A, 3*, 2, 235-241, February 1986.

Attneave, F., "Apparent movement and the what-where connection," *Psychologia, 17*, 108-120, 1974.

Attneave, F. & Block, G., "Apparent movement in tridimensional space," *Perception & Psychophysics, 13*, 2, 301-307, 1973.

Badler, N.I. & Tsotsos, J.K. (eds.), *Motion: Representation and Perception*, Elsevier Science Publishing Co., Inc., 1986.

Baker, C.L., Jr., & Braddick, O.J. "Eccentricity-dependent scaling of the limits for short-range apparent motion perception," *Vision Research, 25*, 6, 803-812, 1985.

Ballard, D.H., "Cortical connections and parallel processing: Structure and function," *The Behavioral and Brain Sciences, 9*, 1, 1986.

Ballard, D.H. & Brown, C.M., *Computer Vision*, Prentice-Hall, 1984.

Banta, A.R. & Breitmeyer, B.G., "Stationary patterns suppress the perception of stroboscopic motion," *Vision Research, 25*, 10, 1501-1505, 1985.

Barlow, H.B., "The twelfth Bartlett memorial lecture: The role of single neurons in the psychology of perception," *Quarterly Journal of Experimental Psychology, 37A*, 121-145, 1985.

Barlow, H.B., "Critical limiting factors in the design of the eye and visual cortex," The Ferrier Lecture, 1980, *Proc. R. Soc. London B 212*, 1-34, 1981.

Barlow, H.B. & Levick, W.R., "The mechanism of directionally selective units in rabbit's retina," *Journal of Physiology, 178*, 477-504, 1965.

Barlow, R. & Verillo, R. *Vision Research*, 1291, 1976.

Barbur, J.L., "Subthreshold addition of real and apparent motion," *Vision Research, 21*, 4, 557-564, 1981.

Baylor, D.A., "Photoreceptor signals and vision," *Investigative Ophthalmology & Visual Science, 28*, 34-49, January 1987.

Börjesson, E. & von Hofsten C., "A vector model for perceived object rotation and translation in space," *Psychol. Research, 38*, 209-230, 1975.

Bonnet, C., "Visual motion detection models: Features and frequency filters," *Perception, 6*, 519-527, 1977.

Bourassa, C.M., "Models for sensation and peception: A selective history," *Human Neurobiology, 5*, 23-26, 1986.

Braddick., O.J., "Low-level and high-level processes in apparent motion," *Phil. Trans. R. Soc. London, B 290*, 137-151, 1980.

Braddick, O.J., "A short-range process in apparent motion," *Vision Research, 14*, 519-527, 1974.

Brandt, T., Dichgans, J. & Koenig, E., "Differential effects of central vs. peripheral vision on egocentric and exocentric motion perception," *Exp. Brain Res., 16*, 476-491, 1973.

Braunstein, M.L., "Perception of rotation in depth: The psychophysical evidence," in *Motion: Representation and Perception*, Elsevier Science Publishing Co., Inc., 1986.

Breitmeyer, B.G. (ed.), *Visual Masking: An Integrative Approach*, Oxford Psychology Series No. 4, Oxford University Press, 1984.

Burbeck, C.A. & Kelly, D.H., "Role of local adaptation in the fading of stabilized images," *J. Optical Society of America*, A, , 2, 216-220, 1984.

Burr, D.C., "Summation of target and mask metacontrast stimuli," *Perception, 13*, 183-192, 1984.

Burr, D.C., Morrone, M.C. & Ross, J., "Seeing objects in motion," *Proceedings, Royal Society of London* B, *227*, 249-265, 1986.

Burr, D.C., Ross, J. & Morrone, M.C., "Smooth and sampled motion," *Vision Research, 26*, 4, 643-652, 1986.

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

Chen, L., "Topological structure in the perception of apparent motion," *Perception, 14*, 197-208, 1985.

Chitty, A.J., Perrett, D.I., Mistlin, A.J. & Harries, M., "Visual cells sensitive to biological motion," submitted for publication.

Chubb, C. & Sperling, G., "Processing stages in non-Fourier motion perception," *Investigative Ophthalmology and Visual Science, 29*, 3, 1988, ARVO Supplement 266.

Chun, H.W., "A representation for temporal sequence and duration in massively parallel networks: Exploiting link interactions," *Proceedings*, AAAI-86, August 1986.

Collewijn, H. & Tamminga, E.P.,"Human fixation and pursuit in normal and open-loop conditions: Effects of central and peripheral retinal targets," *J. Physiol. 379*, 109-129, 1986.

Cooper, P.R. "Structure recognition by connectionist relaxation," *Proc. Canadian Artificial Intelligence Conference CSCSI-88*, Edmonton, Alta., June, 1988, 148-155. Also *Proc., DARPA Image Understanding Workshop,* Cambridge, Mass., April, 1988, 981-994.

Cooper, P.R. & Hollbach, S.C., "Parallel recognition of objects comprosed of pure structure," *Proceedings,* DARDA Image Understanding Workshop, Los Angeles, CA, 1987.

Coren, S., Bradley, D.R., Hoenig, P. & Girgus, J.S., "The effect of smooth tracking and saccadic eye movements on the perception of size: The shrinking circle illusion," *Vision Research, 15*, 49-55, 1975.

Cutting, J.E., "Six tenets for even perception," *Cognition, 10*, 71-78, 1981.

Davidson, M.L., Fox, M-J. & Dick, A.O., "Effect of eye movements on backward masking and perceived location," *Perception & Psychophysics, 14*, 1, 1101-116, 1973.

Dawson, M.R.W., "The cooperative application of multiple natural constraints to the motion correspondence problem," *Proceedings*, Canadian Artificial Intelligence Conf., Edmonton, Alberta, 140-147, June 1988.

De Renzi, E., *Disorders of Space Exploration and Cognition*, John Wiley & Sons, Inc. 1982.

Derrington, A.M. & Henning, G.B., "Errors in direction-of-motion discrimination with complex stimuli," *Vision Research, 27*, 1, 61-75, 1987.

Dick, M., Ullman, S. & Sagi, D., "Parallel and serial processes in motion detection," *Science, 237*, 400-402, 24 July 1987.

Dippé, M.A.Z. & Wold, E.H., "Antialiasing through stochastic sampling," *Comput. Graphics, 19*, 3, 69-78, July 1985.

Ditchburn, R.W., "The function of small saccades," *Vision Research, 20*, 271-272, 1980.

Dursteler, M.R., Wurtz, R.H. & Newsome, W.T., "Directional pursuit deficits following lesions of the foveal representation within the superior temporal sulcus of the macaque monkey," *Jour. Neurophysiology, 57*, 5, 1261-1287, May 1987.

Eckmiller, R., "Neural control of pursuit eye movements," *Physiological Reviews, 67*, 3, 797-857, July 1987.

Elman, J.L., "Finding structure in time," TR 8801, Center for Research in Language, Univ. California, San Diego, 1988.

Farne, M. & Sebellico, A., "Illusory motion induced by rapid displacements of the observer," *Perception, 14*, 393-402, 1985.

Farrell, J.E., "Visual transformations underlying apparent movement," *Perception & Psychophysics, 33*, 1, 85-92, 1983.

Farrell, J.E. & Shepard, R.N., "Shape, orientation, and apparent rotational motion," in *Mental Images and Their Transformations*, R.N. Shepard and L.A. Cooper (eds.), MIT Press, 287-302, 1982.

Feldman, J.A., "Connectionist models and parallelism in high level vision," *Computer Vision, Graphics, and Image Processing, 31*, 178-200, 1985 a.

Feldman, J.A., "Four frames suffice: a provisional model of vision and space, *Behavioral and Brain Sciences, 8*, 265-289, 1985 b.

Feldon, J., Rawlins, J.N.P. & Gray, J.A., "Fornix-fimbria section and the partial reinforcement extinction effect," *Exp. Brain Res., 58*, 435-439, 1985.

Fennema, C. & Thompson, W., "Velocity determination in scenes containing several moving objects," *Computer Graphics and Image Processing, 9*, 301-315, 1979.

Festinger, L., Sedgwick, H.A. & Holtzman, J.D., "Visual perception during smooth pursuit eye movements," *Vision Research, 16*, 1377-1386, 1976.

Finke, R.A., Freyd, J.J. & Shyi, G.C.-W., "Implied velocity and acceleration induce transformations of visual memory," *Jour. Experimental Psychology: General, 115*, 2, 175-188, 1986.

Fleet, D., Hallett, P.E. & Jepson, A.D., "Spatiotemporal inseparability in early visual processing," *Biological Cybernetics, 52*, 153-164, 1985.

Freyd, J.J., "Dynamic mental representations," *Psychological Review, 94*, 4, 427-438, 1987.

Freyd, J.J., "Representing the dynamics of a static form," *Memory & Cognition, 11*, 4, 342-346, 1983.

Freyd, J.J., "The mental representation of movement when static stimuli are viewed," *Perception and Psychophysics, 33*, 6, 575-581, 1983.

Freyd, J.J. & Johnson, J.Q., "Probing the time course of representational momentum," *Jour. Exp. Psychology: Learning, Memory and Cognition, 13*, 2, 259-268, 1987.

Georgopoulos, A.P., "On reaching," *Ann. Rev. Neurosci., 9*, 147-170, 1986.

Gibson, J.J., "What gives rise to the perception of motion?" *Psychological Review, 75*, 335-346, 1968.

Goddard, N.H., "The representation and use of visual motion," Ph.D. dissertation, Computer Science Dept., Univ.Rochester, forthcoming.

Green, D.G., "The search for the site of visual adaptation," *Vision Research, 26*, 9, 1417-1430, 1986.

Green, M., "Inhibition and facilitation of apparent motion by real motion," *Vision Research, 23*, 9, 861-865, 1983.

Green, M. & Odom, J.V., "Correspondence matching in apparent motion: Evidence for three-dimensional spatial representation," *Science, 233*, 1427-1429, 26 September 1986.

Green, M. & von Grunau, M., "Real and apparent motion: One mechanism or two?" in *Motion: Representation and Perception*, Elsevier Science Publishing Co., Inc., 1986.

Gregory, R.K., "Movement nulling: For heterochromatic photometry and isolating channels for 'real' and 'apparent' motion," *Perception, 14*, 193-196, 1985.

Haber, R.N., & Hershenson, M., *The Psychology of Visual Perception*, Holt, Rinehart & Winston, 1980.

Hansen, R.M. & Skavenski, A.A., "Accuracy of eye position information for motor control," *Vision Research, 17*, 919-926, 1977.

Hayhoe, M.M., Benimoff, N.I. & Hood, D.C., "The time course of multiplicative and subtractive adaptation processes," *Vision Research*, in press.

Hildreth, E.C. & Koch, C. "The analysis of visual motion: From computational theory to neuronal mechanisms," *Ann. Rev. Neuroscience, 10*, 477-533, 1987.

Hoenkamp, E., "Perceptual cues that determine the labeling of human gait," *Journal of Human Movement Studies, 4*, 59-69, 1978.

Hoffman, D.D. & Flinchbaugh, B.E., "The interpretation of biological motion," *Biological Cybernetics, 42*, 195-204, 1982.

Hogben, J.H. & DiLollo, V., "Practice reduces suppression in metacontrast and in apparent motion," *Perception & Psychophysics, 35*, 5, 441-445, 1984.

Holtzman, J.D., Sedgwick, H.A. & Festinger, L., "Interaction of perceptually monitored and unmonitored efferent commands for smooth pursuit eye movements," *Vision Research, 18*, 1545-1555, 1978.

Hrechanyk, L.M. & Ballard, D.H., "A connectionist model of form perception," *Proceedings*, IEEE Computer Vision Workshop, Rindge, NH, August 1982, 44-52.

Hubel, D.H., *Eye, Brain, and Vision*, Scientific Americal Library, 1988.

Hubel, D.H. & Wiesel, T.N., "Functional architecture of macaque monkey visual cortex," *Proc. R. Soc. London B 172*, 563-584, 1977.

Hutchison, J., Koch, C., Luo, J. & Mead, C., "Computing motion using analog and binary resistive networks," *Computer, 21*, 3, 52-64, 1988.

Ikeda, M., "Temporal impulse response," *Vision Research, 26*, 9, 1431-1442, 1986.

Irwin, D.E. & Brown, J.S., "Visual masking and visual integration across saccadic eye movements," *Journ. Experimental Psychology: General*, in press.

Iwai, E., "Neuropsychological basis of pattern vision in macaque monkeys," *Vision Research, 25*, 3, 425-439, 1985.

Jacobs, G.H., "Cones and opponency," *Vision Research, 26*, 9, 1533-1542, 1986.

Jagacinski, R.J., Johnson, W.W. & Miller, R.A., "Quantifying the cognitive trajectories of extrapolated movements," *Jour. Experimental Psychology: Human Perception and Performance, 9*, 1, 43-57, 1983.

Jenkin, M., "Tracking three-dimensional moving light displays," in *Motion: Representation and Perception*, Proceedings, ACM SIGGRAPH/SIGART Interdisc. Workshop on Motion: Representation and Perception, Toronto, Ontario, North-Holland, 171-175, 1983.

Johansson, G., "Spatio-temporal differentiation and integration in visual motion perception," *Psychological Research, 38*, 379-393, 1976.

Johansson, G., "Visual motion perception," *Scientific American, 232*, 6, 76-88, 1975.

Johannsen, G., "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics, 14*, 201-211, 1973.

Kawano, K. & Miles, F.A. "Short-latency ocular following responses of monkey. II: Dependence on a prior saccadic eye movement," *Jour. Neurophysiology, 56*, 5, 1355-1380, November 1986.

Kawano, K., Sasaki, M. & Yamashita, M., "Response properties of neurons in posterior parietal cortex of monkey during visual-vestibular sitimulation. I. Visual tracking neurons," *Jour. Neurophysiology, 551*, 2, 340-351, 1984.

Kelly, D.H., "Visual processing of moving stimuli," *Journal of the Optical Society of America* A, *2*, 2, 216-225, 1985.

Kelly, M.H. & Freyd, J.J., "Explorations of representational momentum," *Cognitive Psychology, 19*, 369-401, 1987.

Kleinfeld, D., "Sequential state generation by model neural networks," *PNAS*, 83, 1987.

Klopfer, D.S. & Cooper, L.A., "Using apparent motion to measure the structure of perceived space," presented at the 26th annual meeting, Psychonomic Society, Boston, MA, 1985.

Koch, C., Poggio, T. & Torre, V., "Computations in the vertebrate retina: Gain enhancement, differentiation and motion discrimination," *Trends in NeuroSciences, 9*, 5, 204-210, 1986.

Kolers, P.A., "Motion from continuous or discontinuous arrangements," in *Motion: Representation and Perception*, Proceedings, ACM SIGGRAPH/SIGART

Interdisc. Workshop on Motion: Representation and Perception, Toronto, Ontario, North-Holland, 227-241, 1983.

Kolers, P.A., *Aspects of Motion Perception*, Pergamon Press, 1972.

Kolers, P.A. & Pomerantz, J.R., "Figural change in apparent motion," *Journal of Experimental Psychology, 87*, 99-108, 1971.

Kolers, P.A. & von Grunau, M., "Shape and color in apparent motion," *Vision Research, 16*, 329-335, 1976.

Kosslyn, S.M., "Seeing and imagining in the cerebral hemispheres: A computational approach," *Psychological Review, 94*, 2, 148-175, 1987.

Kosslyn, S.M., Flynn, R.A. & Amsterdam, J.B., "Components of high-level vision: A cognitive neuroscience analysis," submitted for publication.

Kowler, E. & Steinman, R.M. "Small saccades serve no useful purpose: reply to a letter by R.W. Ditchburn," *Vision Research, 20*, 273-276, 1980.

Knudsen, E.R., de Lac, S. & Esterly, S.D., "Computational maps in the brain," *Ann. Rev. Neurosci, 10*, 41-65, 1987.

Kozlowski, L.T. & Cutting, J.E., "Recognizing the sex of walker from dynamic point-light displays," *Perception and Psychophysics, 21*, 6, 575-580,

Lawton, D., Rieger, J. and Steenstrup, M., "Computational techniques in motion processing," in M.A. Arbib and A.R. Hanson (eds.), *Vision, Brain and Cooparative Computation*, Bradford Books/MIT Press, 1987.

Lee, D.N. & Lishman, J.R., "Visual control of locomotion," *Scand. J. Psych., 18*, 224-230, 1977.

Lee, D.N., Young, D.S., Reddish, P.E., Lough, S. & Clayton, T.M.H., "Visual timing in hitting an accelerating ball," *Quarterly Journal of Expeimental Psychology, 35A*, 333-346, 1983.

Lennie, P., "Parallel visual pathways: A review," *Vision Research, 20*, 561-595, 1980.

Levick, W.R. & Dvorak, D.R.,"The retina -- from molecules to networks," *Trends in NeuroSciences, 9*, 5, 181-185, May 1986.

Linsker, R., "Self-organization in a perceptual network," *Computer, 21*, 3, 105-117, 1988.

Lisberger, S.G., Morris, E.J. & Tychsen, L., "Visual motion processing and sensory-motor integration for smooth pursuit eye movements," *Ann Rev. Neurosci, 10*, 97-129, 1987.

Longuet-Higgins, H.C. & Prazdny, K., "The interpretation of a moving retinal image," *Proceedings, Royal Society of London B, 208*, 385-397, 1980.

Lowe, D.G., "The viewpoint consistency constraint," *International Journal of Computer Vision, 1*, 57-72, 1987.

Mack, A. & Bachant, J., "Perceived movement of the afterimage during eye movements," *Perception and Psychophysics, 6*, 6A, 379-384, 1969.

Mack, A., Fendrich, R. & Wong, E., "Is perceived motion a stimulus for smooth pursuit," *Vision Research, 22*, 77-88, 1982.

Macko, K.A., Jarvis, C.D., Kennedy, C., Miyaoka, M., Shinohara, M., Sokoloff, L. & Mishkin, M., "Mapping the primate visual system with [2-^{14}C] Deoxyglucose," *Science, 218*, 394-397, 22 October 1982.

Madden, B.C., "Apparent motion, real effects," TR 247, Computer Science Dept., Univ. Rochester, to appear, 1988a.

Madden, B.C., "Space, time and apparent motion," TR 246, Computer Science Dept., Univ. Rochester, to appear, 1988b.

Madden, B.C., "A theory of spatial acuity," Ph.D. dissertation, Psychology Dept., Univ. Rochester, 1985.

Marr, D., *Vision*, Freeman Publishing Company, 1982.

Marr, D. & Ullman, S., "Directional selectivity and its use in early visual processing," *Proc. R. Soc. Lond., B 211*, 151-180, 1981.

Mather, G. & Anstis, S., "Motion perception: Second thoughts on the correspondence problem," in *Motion: Representation and Perception*, Proceedings, ACM SIGGRAPH/SIGART Interdisc. Workshop on Motion: Representation and Perception, Toronto, Ontario, North-Holland, 63-78, 1983.

Matin, L., Pola, J., Matin, E. & Picoult, E., "Vernier discrimination with sequentially-flashed lines: Roles of eye movements, retinal offsets and short-term meory," *Vision Research, 21*, 647-656, 1981.

Maunsell, J.H.R., "Physiological evidence for two visual subsystems," in *Matters of Intelligence*, L.M. Vaina (ed.), D. Reidel Publishing Company, 59-87, 1987.

Maunsell, J.H.R. & Newsome, W.T., "Visual processing in monkey extrastriate cortex," *Annual Review of Neuroscience, 10*, 363-401, 1987.

McLeod, P. Driver, J. & Crisp, J., "Visual search for a conjunction of movement and form is parallel," *Nature, 332*, 154-155, 10 March 1988.

McKee, S.P., "A local mechanism for differential velocity detection," *Vision Research, 21*, 491-500, 1981.

McKee, S.P. & Nakayama, K., "The detection of motion in the peripheral visual field," *Vision Research, 24*, 1, 25-32, 1984.

McKee, S.P. & Welch, L, "Sequential recruitment in the discrimination of velocity," *Jour. Optical. Soc. America., A, 2*, 2, 243-251, February 1985.

McKee S.P. & Welch, L., "Is there constancy for velocity?" unpublished manuscript, 1988.

McKee, S.P., Silverman, G.H. & Nakayama, K., "Precise velocity discrimination despite random variations in temporal frequency and contrast," *Vision Research, 26*, 4, 609-619, 1986.

Mikami, A., Newsome, W.T. & Wurtz, R.H. "Motion selectivity in macaque visual cortex. II. Spatiotemporal range of directional interactions in MT and V1," *Jour. of Neurophysiology, 55*, 6, 1328-1339, June 1986.

Miles, F.A. & Kawano, K., "Short-latency ocular following responses of monkey. III: Plasticity," *Jour. Neurophysiology, 56*, 5, 1381-1396, November 1986.

Miles, F.A., Kawano, K. & Optican, L.M., "Short-latency ocular following responses of monkey. I: Dependence on temporospatial properties of visual input," *Jour. Neurophysiology, 56*, 5, 1321-1354, November 1986.

Mitrani, L., Radil-Weiss, T., Yakimoff, N., Mateeff, St. & Bozkov, V., "Deterioration of vision due to contour shift over the retina during eye movements," *Vision Research, 15*, 877-878, 1975.

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

Morgan, M.J. & Watt, R.J., "Spatial frequency interference effects and interpolation in vernier accuity," *Vision Research, 24*, 1911-1919, 1984.

Morgan, M.J. & Watt, R.J., "Effect of motion sweep duration and number of stations upon interpolation in discontinuous motion," *Vision Research, 22*, 1277-1284, 1982.

Motter, B.C., Steinmetz, M.A., Duffy, C.J. & Mountcastle, V.B., "Functional properties of parietal visual neurons: Mechanisms of directionality along a single axis, *Journal of Neuroscience, 7*, no. 1, January 1987.

Murphy, P.C. & Sillito, A.M., "Corticofugal feedback influences the generation of length tuning in the visual pathway," *Nature, 329*, 727-729, 22 Oct. 1987.

Mutch, K., Smith, I.M. & Yonas, A., "The effect of two-dimensional and three-dimensional distance on apparent motion," *Perception, 12*, 305-312, 1983.

Nakayama, K., "Biological image motion processing: A review," *Vision Research, 25*, 5, 625-660, 1985.

Nakayama, K. & Silverman, G., *Nature, 320*, 264-265, 1986.

Neumann, B., "Optical flow," in *Motion: Representation and Perception*, N.I. Badler & J.K. Tsotsos (eds.), Elsevier Science Publishing Co., Inc., 1983.

Newsome, W.T., Gizzi, M.S. & Movshon, J.A., "Spatial and temporal properties of neurons in macaque MT," *Investigative Ophthalmology and Visual Science, 24*, 106, 1983.

Newsome, W.T., Mikami, A. & Wurtz, R.H.,"Motion selectivity in macaque visual cortex. III. Psychophysics and physiology of apparent motion," *Jour. of Neurophysiology, 55*, 6, 1340-1351, June 1986.

O'Connell, B.G. & Gerard, A.B., "Scripts and scraps: The development of sequential understanding," *Child Development, 56*, 671-681, 1985.

Olson, T.J., "A two-stage model of motion understanding," Ph.D. dissertation, Computer Science Dept., Univ. Rochester, forthcoming, 1988.

Pantle, A., "Stroboscopic movement based upon global information in successively presented visual patterns," *J. Opt. Soc. Am., 63*, 1280 A, 1973.

Pantle, A. & Picciano, L., "A multistable movement display: Evidence for two separate motion systems in human vision," *Science 193*, 500-502, 1976.

Perrett, D.I., Harries, M., Mistlin, A.J. & Chitty, A.J., "Three stages in the classification of body movements by visual neurons," in Proceedings, Intern'l. Symp. on Images and Understanding, H. Barlow, C. Blakemore, M. Weston-Smith (eds.), in press.

Perrett, D.I., Mistlin, A.J. & Chitty, A.J., "Visual neurones responsive to faces," *Trends in NeuroSciences, 10*, 9, 358-364, 1987.

Perrett, D.I., Smith, P.A.J., Mistlin, A.J., Chitty, A.J., Head, A.S., Potter, D.D., Broennimann, R., Milner, A.D. & Jeeves, M.A. "Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report," *Behavioural Brain Research, 16*, 153-170, 1985.

Petersik, J.T., "The effects of spatial and temporal factors on the perception of stroboscopic rotation simulations," *Perception, 9*, 271-283, 1980.

Pettigrew, J.D., Sanderson, K.J. & Levick, W.R. (eds.), *Visual Neuroscience*, Cambridge University Press, 1986.

Plaut, D.C., "Visual recognition of simple objects by a connection network," TR 143, Computer Science Dept., Univ. Rochester, 1984.

Poggio, T., Torre, V. & Koch, C., "Computational vision and regularization theory," *Nature, 317*, 314-319, 1985.

Post, R.B. & Leibowitz, H.W., "A revised analysis of the role of efference in motion perception," *Perception, 14*, 631-643, 1985.

Priest, H.F. & Cutting, J.E., "Visual flow and direction of locomotion," *Science, 227*, 1063-1064, 1 March 1985.

Prazdny, K., "What variables control (long-range) apparent motion?" *Perception, 15*, 37-40, 1986a.

Prazdny, K., "Three-dimensional structure from long-range apparent motion," *Perception, 15*, 619-625, 1986b.

Prazdny, K., " 'Capture of stereopsis by illusory contours,' " *Nature, 324*, 393-394, 27 November 1986c.

Ramachandran, V.S., "Interaction between colour and motion in human vision," *Nature, 328*, 645-647, 13 August 1987.

Ramachandran, V.S., "Apparent motion of subjective surfaces," *Perception, 14*, 127-134, 1985.

Ramachandran, V.S. & Anstis, S., "Figure-ground segregation modulates apparent motion," *Vision Research, 26*, 12, 1969-1975, 1986.

Ramachandran, V.S. & Anstis, S., "Extrapolation of motion path in human visual perception," *Vision Research, 23*, 83-85, 1983.

Ramachandran, V.S. & Cavanagh, P., "Motion capture anisotropy," *Vision Research, 27*, 1, 97-106, 1987.

Ramachandran, V.S. & Cronin-Golomb, A. & Myers, J.J., "Perception of apparent motion by commissurotomy patients," *Nature, 320*, 358-359, 27 March 1986.

Ramachandran, V.S., Ginsburg, A.P. & Anstis, S.M., "Low spatial frequencies dominate apparent motion," *Perception, 12*, 1983.

Ramachandran, V.S., Inada, V., & Kiama, G., "Perception of illusory occlusion in apparent motion," *Vision Research, 26*, 10, 1741-1749, 1986.

Rashid, R.F., "Towards a system for the interpretation of moving light displays," *IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-2*, no. 6, 574-581, November 1980.

Regan, D. & Beverley, K.I., "How do we avoid confounding the direction we are looking and the direction we are moving?" *Science, 215*, 194-196, 8 January 1982.

Reichardt, W., "Autocorrelation, a principle for the evaluation of sensory information by the central nervous system," in *Sensory Communication*, W.A. Rosenblith (ed.), John Wiley, New York, 1961.

Richter, J. & Ullman, S., "A model for the temporal organization of X- and Y-type receptive fields in the primate retina," *Biological Cybernetics, 43*, 127-145, 1982.

Robinson, D.L., Goldberg, M.E. & Stanton, G.B., "Parietal association cortex in the primate: Sensory mechanisms and behavioral modulations," *Jour. Neurophysiology, 41*, 4, 910-932, 1978.

Runeson, S. & Frykholm, G., "Kinematic specification of dynamics as an information basis for person-and-action perception: Expectation, gender recognition, and deceptive intention," *Jour. Experimental Psychology: General, 112*, 4, 585-615, 1983.

Sagi, D. & Julesz, B., " 'Where' and 'what' in vision," *Science, 228*, 1217-1219, 7 June 1985.

Sakata, H., Shibutani, H. & Kawano, K., "Functional properties of visual tracking neurons in posterior parietal association cortex of the monkey," *Jour. Neurophysiology, 49*, 6, 1364-1380, 1983.

Sakata, H., Shibutani, H. & Kawano, K., "Parietal neurons with dual sensitivity to real and induced movements of visual target," *Neuroscience Letters, 9*, 165-169, 1978.

Sakata, H., Shibutani, H., Kawano, K. & Harrington, T.L., "Neural mechanisms of space vision in the parietal association cortex of the monkey," *Vision Research, 25*, 3, 453-463, 1985.

Schiller, P.H., "The central visual system," *Vision Research, 26*, 9, 1351-1388, 1986.

Shapley, R. & Perry, V.H., "Cat and monkey retinal ganglion cells and their visual functional roles," *Trends in NeuroSciences, 9*, 5, 229-235, 1986.

Shastri, L., *Semantic Networks*, Pitman, 1988.

Shepard, R.N., "Ecological constraints on internal representations," expanded version, 3rd James J. Gibson Memorial Lecture, Cornell University, October 1983.

Shepard, R.N. and Zare, S.L., "Path-guided apparent motion," *Science, 220*, 632-634, 1982.

Sperling, H.G., "Spectral sensitivity, intense spectral light studies, and the color receptor mosaic of primates," *Vision Research, 26*, 9, 1557-1572, 1986.

Steinman, R.M., "Eye movement," *Vision Research, 26*, 9, 1389-1400, 1986.

Stoper, A. & Mansfield, J.G., "Metacontrast and paracontrast suppression of a contourless area," *Vision Research, 18*, 1669-1674, 1978.

Strelow, E.R., "What is needed for a theory of mobility: Direct perception and cognitive maps -- lessons from the blind," *Psychology Review, 92*, 2, 226-248, 1985.

Sumi, S., "Upside-down presentation of the Johansson moving light-spot pattern," *Perception, 13*, 283-286, 1984.

Thompson, P., "The coding of velocity of movement in the human visual system," *Vision Research, 24*, 1983.

Thompson, P., "Discrimination of moving gratings at and above detection threshold," *Vision Research, 23*, no. 12, 1533-1538, 1983.

Todd, J.T., "Perception of gait," *Jour. Experimental Psychology: Human Perception and Performance, 9*, 1, 31-42, 1983.

Todd, J.T., "Visual information about rigid and non-rigid motion: A geometric analysis," *Journal of Experimental Psychology: Human Perception and Performance, 8*, 238-252, 1982.

Todd, J.T., "Visual information about moving objects," *Journal of Experimental Psychology: Human Perception and Performance, 7*, 4, 794-810, 1981.

Torre, V. & Poggio, T., "A synaptic mechanism possibly underlying directional selectivity to motion," *Proceedings, Royal Society of London B, 202*, 409-416, 1978.

Trends in NeuroSciences, Special issue on information processing in the retina, *8*, 5, 181-240, May 1986.

Ullman, S., *The Interpretation of Visual Motion*, MIT Press, 1979.

Ungerleider, L.G. & Brody, B.A., "Extrapersonal spatial orientation: The role of posterior parietal, anterior frontal, and inferotemporal cortex," *Experimental Neurology, 56*, 265-280, 1977.

Ungerleider, L.G. & Mishkin, M., "Two cortical visual systems," in *Analysis of Visual Behavior*, D.J. Ingle, M.A. Goodale, & R.J.W. Mansfiled (eds.), MIT Press, 1982.

van Doorn, A.J., van de Grind, W.A. & Koenderink, J.J. (eds.) *Limits in Perception*, VNU Science Press, Utrecht, 1984.

Van Essen, D.C. & Felleman, D., personal communication, 1988.

van Santen, J.P.H. & Sperling, G., "Elaborated Reichardt detectors," *Journal of the Optical Society of America A, 2*, 2, 1985.

Verri, A. & Poggio, T., "Qualitative information in the optical flow," *Proceedings*, DARPA Image Understanding Workshop, Los Angeles, CA., 825-834, February 1987.

Volkmann, F.C., "Human visual suppression," *Vision Research, 26*, 9, 1417-1430, 1986.

von Grünau, M., "The involvement of illusory contours in stroboscopic motion," *Perception and Psychophysics, 25*, 3, 205-208, 1979.

von Grünau, M., "Form information is necessary for the perception of motion," *Vision Research, 19*, 839-341, 1979.

Wallach, H., "Perceiving a stable environment when one moves," *Ann. Rev. Psychol., 38*, 1-27, 1987.

Wallach, H., "Keynote Address: How human perception deals with motion," in *Motion: Representation and Perception*, Proceedings, ACM SIGGRAPH/SIGART Interdisc. Workshop on Motion: Representation and Perception, Toronto, Ontario, North-Holland, 1-19, 1983.

Watson, A.B., "Apparent motion occurs only between similar spatial frequencies," *Vision Research, 26*, 10, 1727-1730, 1986.

Watson, A.B. & Ahumada, A.J., Jr.,"Model of human visual-motion sensing," *Journal of the Optical Society of America A, 2*, 2, 1985.

Webb, J.A., "Static analysis of moving jointed objects," *Proceedings*, Conf. of Amer. Assoc. for Artificial Intelligence 1980, 35-37, 1980.

Welch, L. & McKee, S.P., "Colliding targets: Evidence for spatial localization within the motion system," *Vision Research, 25*, 12, 1901-1910, 1985.

Weller, R.E. & Kaas, J.H., "Subdivisions and connections of inferior temporal cortex in owl monkeys," *Jour. Comparative Neurology, 256*, 137-172, 1987.

Wertheim, A.H., Wagenaar, W.A. and Leibowitz, H.W., *Tutorials on Motion Perception*, Plenum Press, 1982.

Westheimer, G., "Physiological optics during the first quarter century of *Vision Research*," *Vision Research, 26*, 9, 1515-1522, 1986.

Westheimer, G. & McKee, S.P., "The perception of temporal order," *Vision Research*, 1977.

Williams, D.R., "Seeing through the photoreceptor mosaic," *Trends in NeuroSciences, 9*, 5, 193-197. 1986.

Williams, M.C. & Weisstein, N., "The effect of perceived depth and connectedness on metacontrast functions," *Vision Research, 24*, 10, 1279-1288, 1984.

Williams, D., Phillips, G. & Sekuler, R., "Hysteresis in the perception of motion direction as evidence for neural cooperativity," *Nature, 324*, 253-255, 20 November 1986.

Wilson, H., "A model for direction selectivity in threshold motion perception," *Biological Cybernetics*, 1985.

Winterson, B.J. & Collewijn, H., "Microsaccades during finely-guided visuomotor tasks," *Vision Research, 16*, 1387-1389, 1976.

Wurtz, R.H. & Mohler, C.W., "Enhancement of visual responses in monkey striate cortex and frontal eye fields," *Jour. Neurophysiology, 39*, 4, 766-772, 1976.

Yates, J., "The content of awareness is a model of the world," *Psychological Review, 92*, 2, 249-284, 1985.

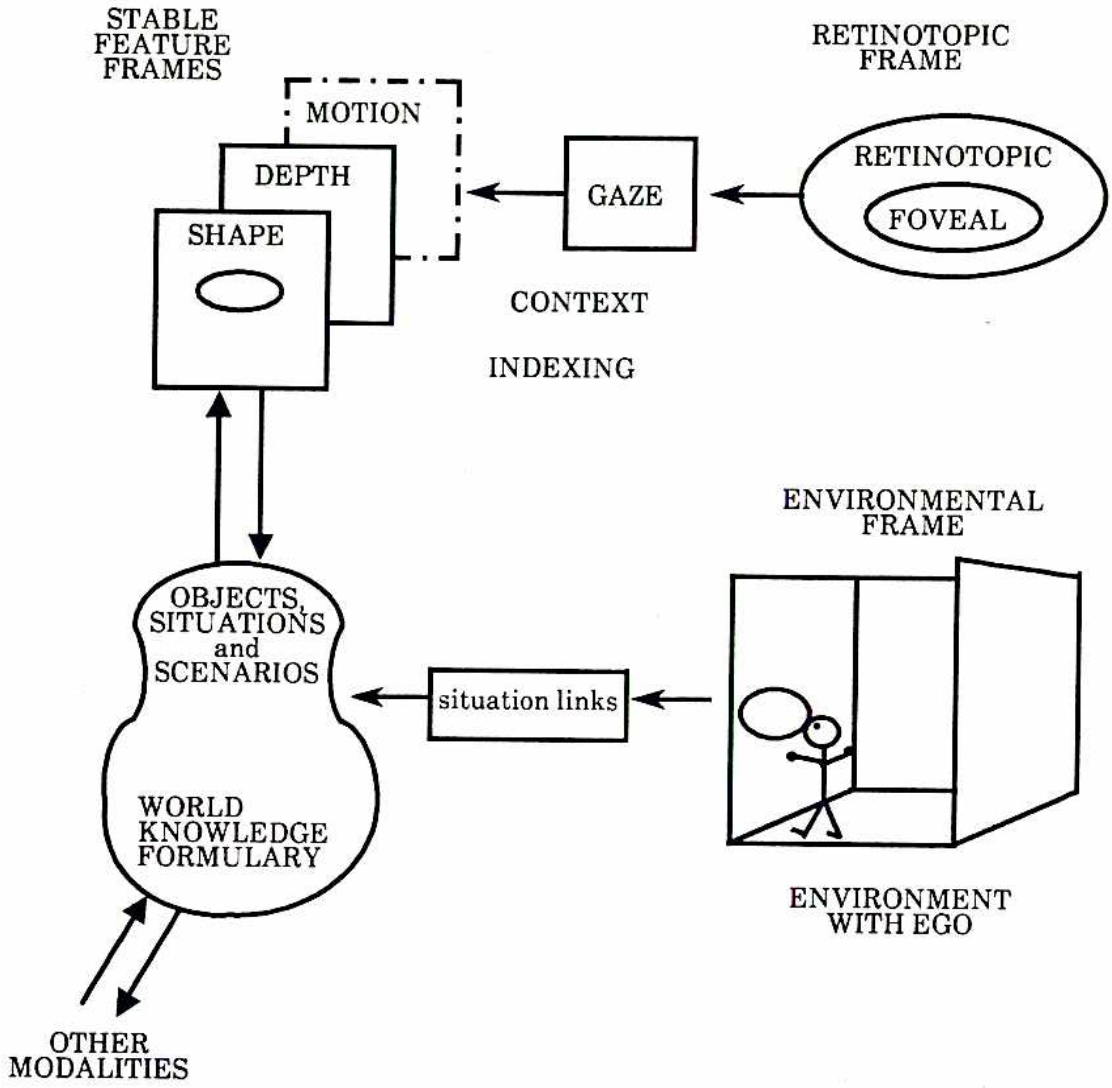


Figure 1: Four Frames

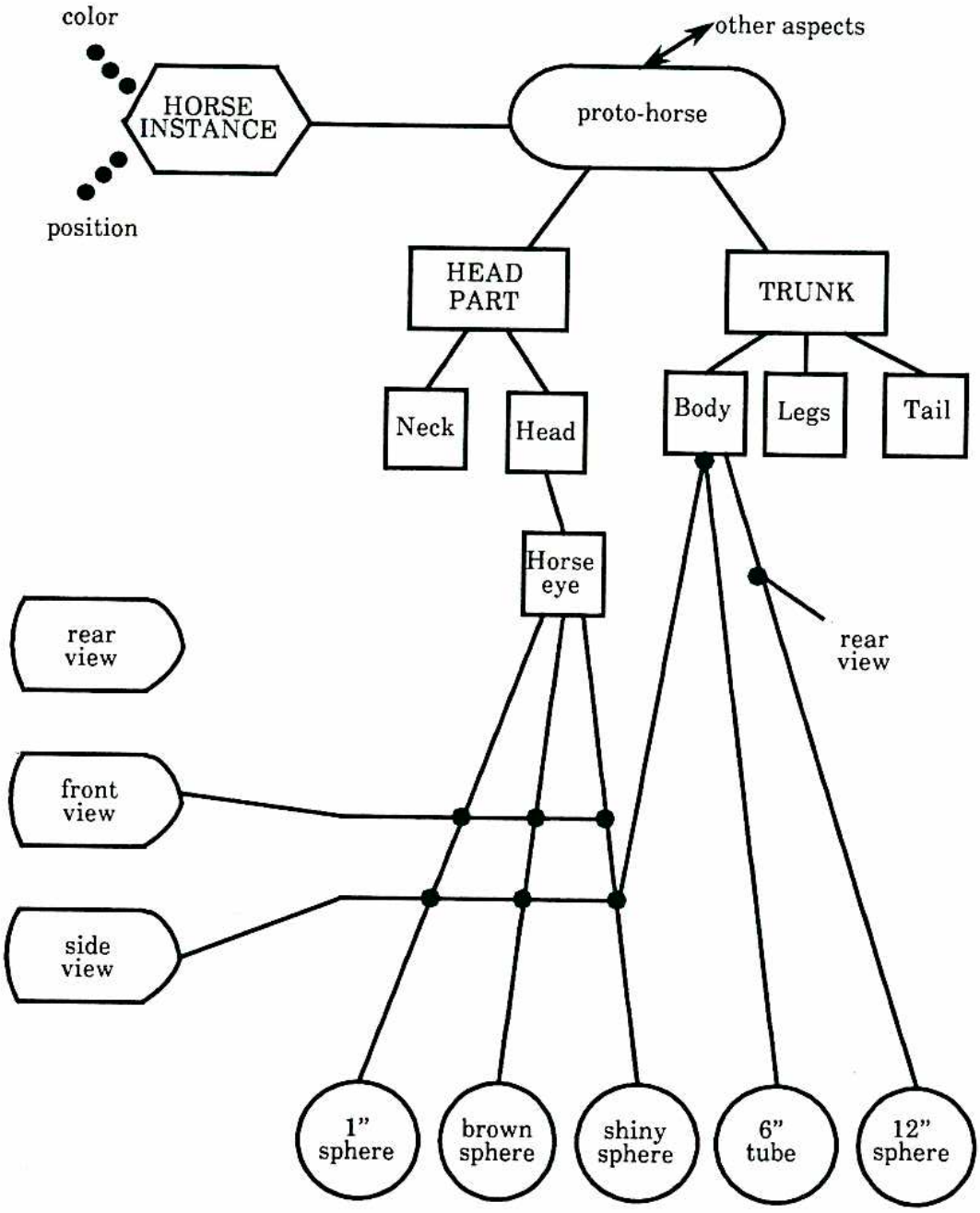

color
position
HORSE INSTANCE
other aspects
proto-horse
HEAD PART
TRUNK
Neck
Head
Body
Legs
Tail
Horse eye
rear view
front view
side view
rear view
1" sphere
brown sphere
shiny sphere
6" tube
12" sphere

Figure 2

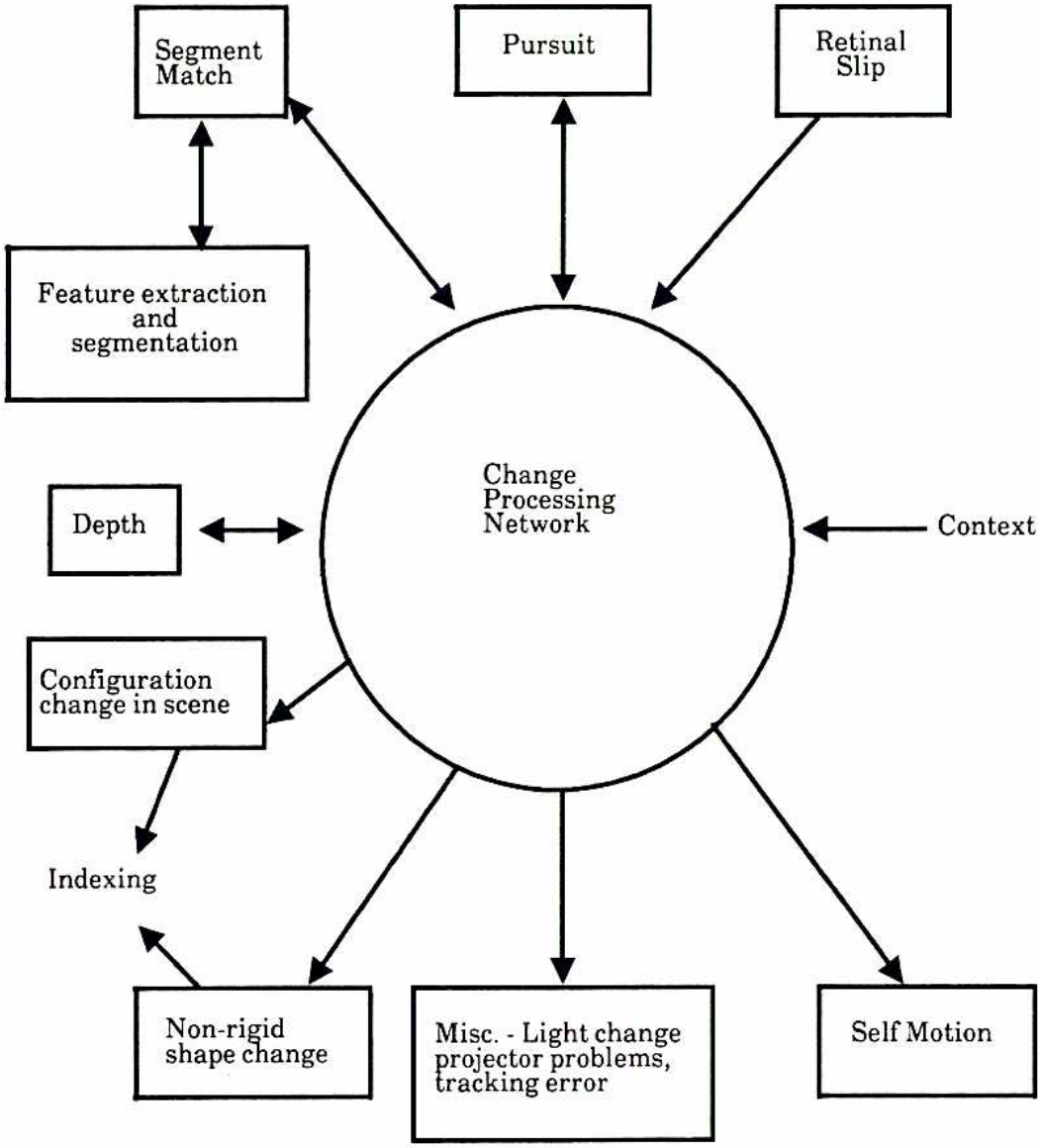

Segment Match
Pursuit
Retinal Slip
Feature extraction and segmentation
Change Processing Network
Depth
Context
Configuration change in scene
Indexing
Non-rigid shape change
Misc. - Light change projector problems, tracking error
Self Motion

Figure 3

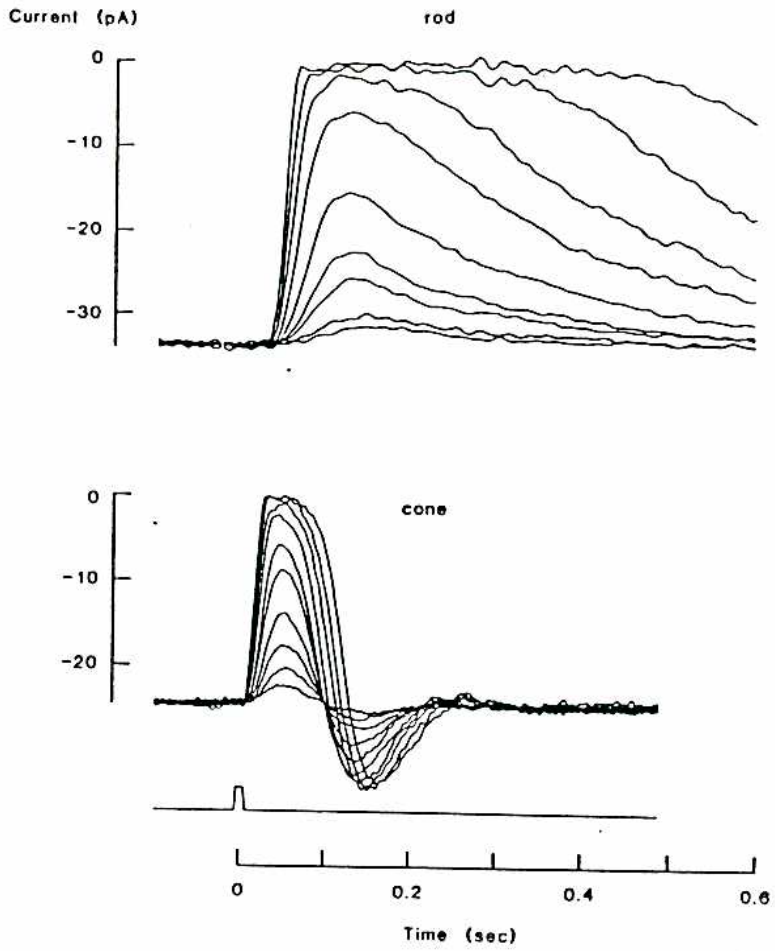


Figure 4 (after Baylor): Photocurrent families recorded from a rod and red cone of the monkey *Macaca fascicularis*. Flash monitor trace below. The ordinate is the membrane current collected from the outer segment by the suction electrode. Flash strengths were increased by factors of 2, and the responses grew to the saturating amplitude, at which the dark current was completely shut off. Responses from the rod are slower and more sensitive than those from the cone, and lack the undershoot present in the cone responses. For the rod, the flashes were expected to cause between 2.9 and 860 photoisomerizations, while for the cone the corresponding figures were 190 and 36,000 photoisomerizations. Some flash responses have been averaged from multiple trials to reduce noise.

Figure 5 (after Burr & Ross): A photograph taken with 125-msec shutter speed. If the temporal summation observed by Burr (1981) acted like a frame store, we would expect motion blur of this magnitude.

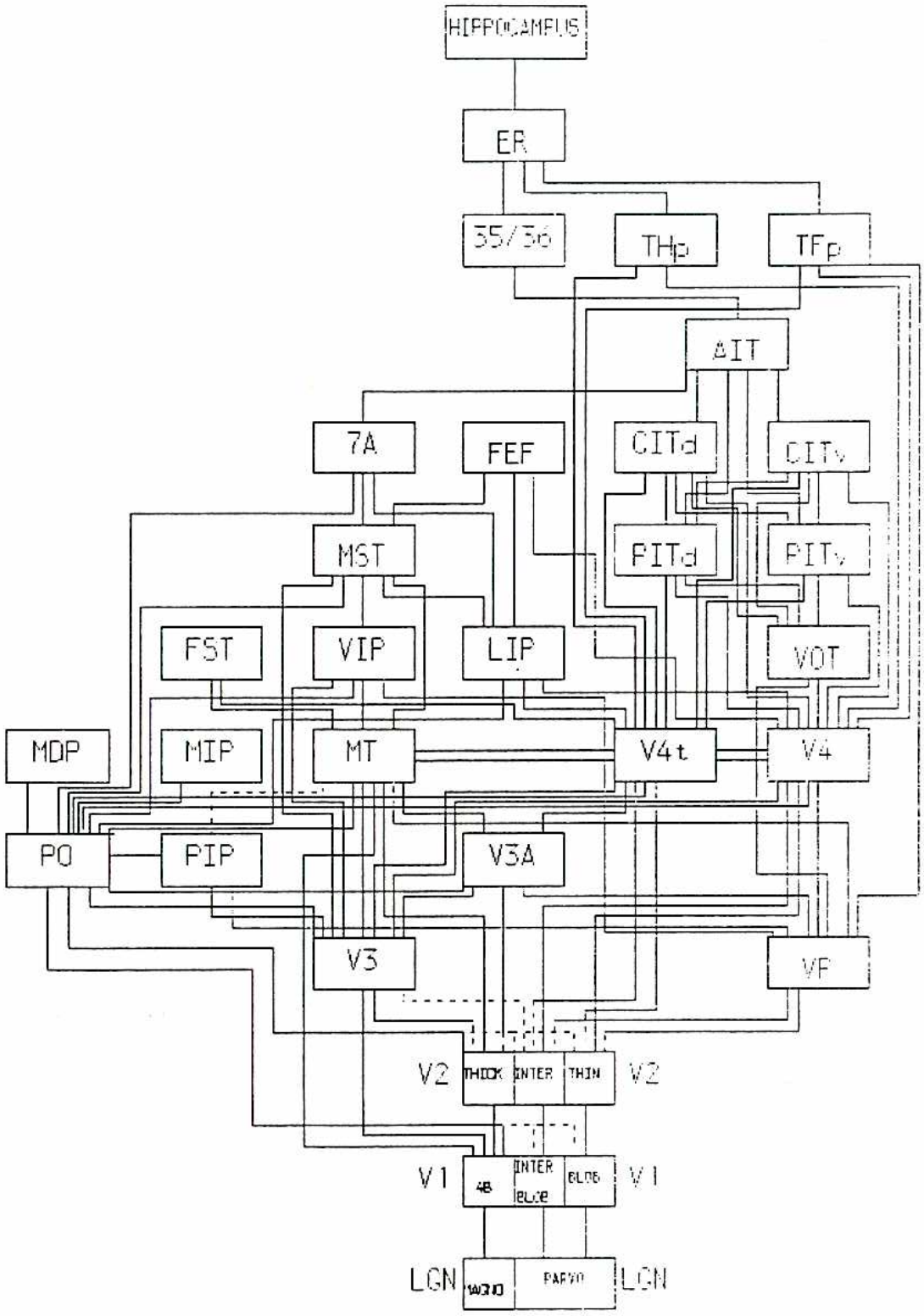


Figure 6 (after Felleman and Van Essen)

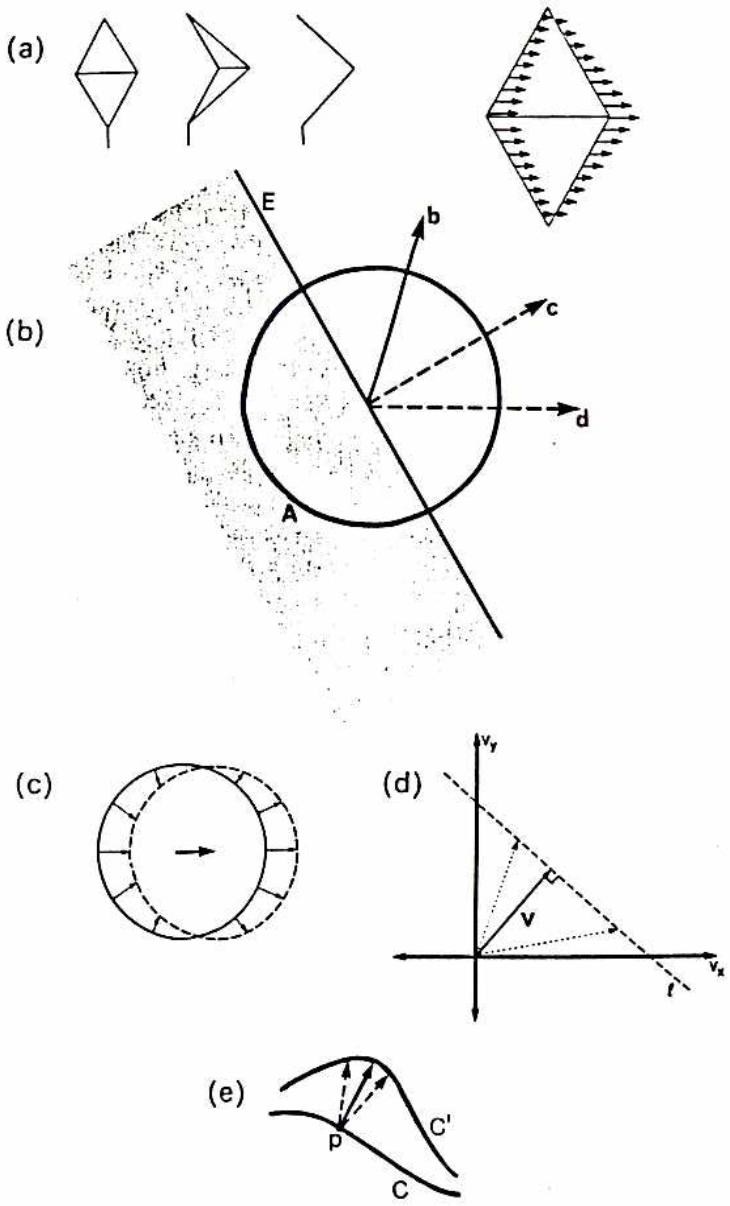


Figure 7 (after Hildreth & Koch): The aperture problem in motion measurement. (a) On the *left* are three views of a wire-frame object undergoing rotation arund a central vertical axis. On the *right*, the *arrows* along the contours of the object represent the instantaneous velocity field at one position in the object's trajectory. For simplicity, an orthographic projection is used. (b) An operation that views the moving edge E through the local aperture A can compute only the component of motion c in the direction perpendicular to the orientation of the edge. The true motion of the edge is ambiguous. (c) The circle undergoes pure translation to the right: the *arrows* represent the perpendicular components of velocity that can be measured from the changing image. (d) The vector v represents the perpendicular component of velocity at some location in the image. The true velocity at that location must project to the line l perpendicular to v; examples are shown with *dotted arrows*. (e) The curve C rotates, translates, and deforms over time to yield the curve C′. The velocity of the point p is ambiguous.

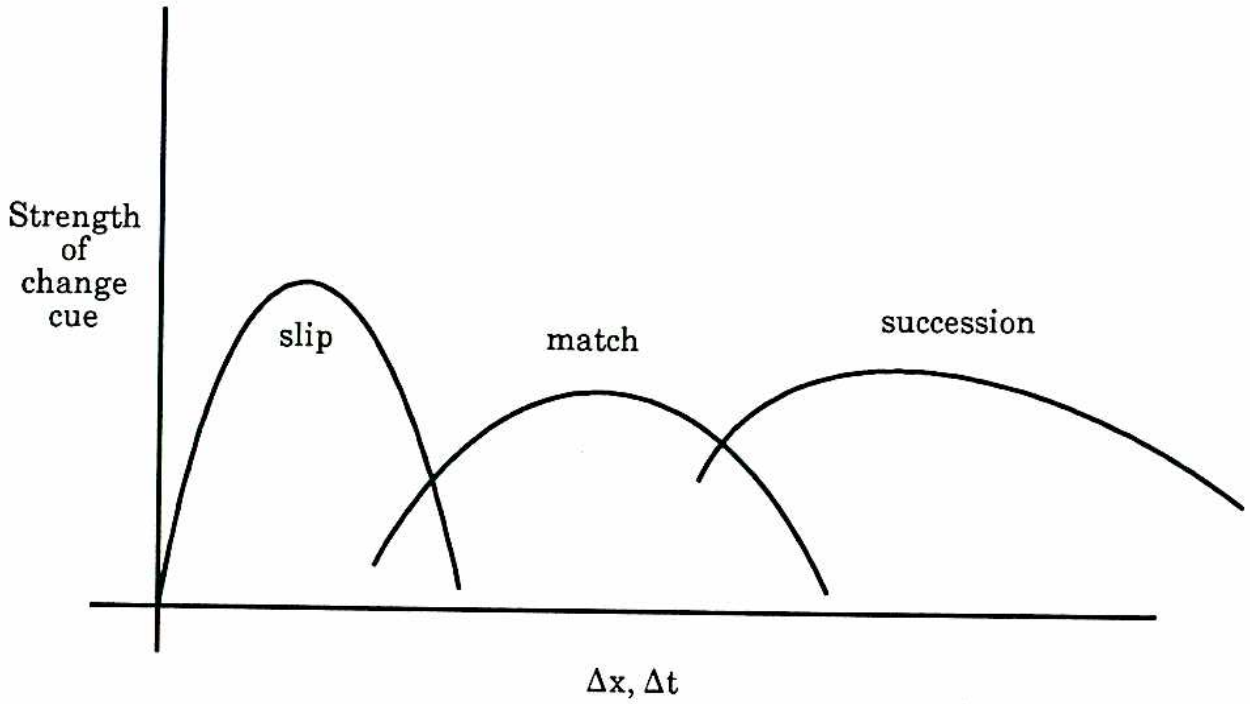


Figure 8

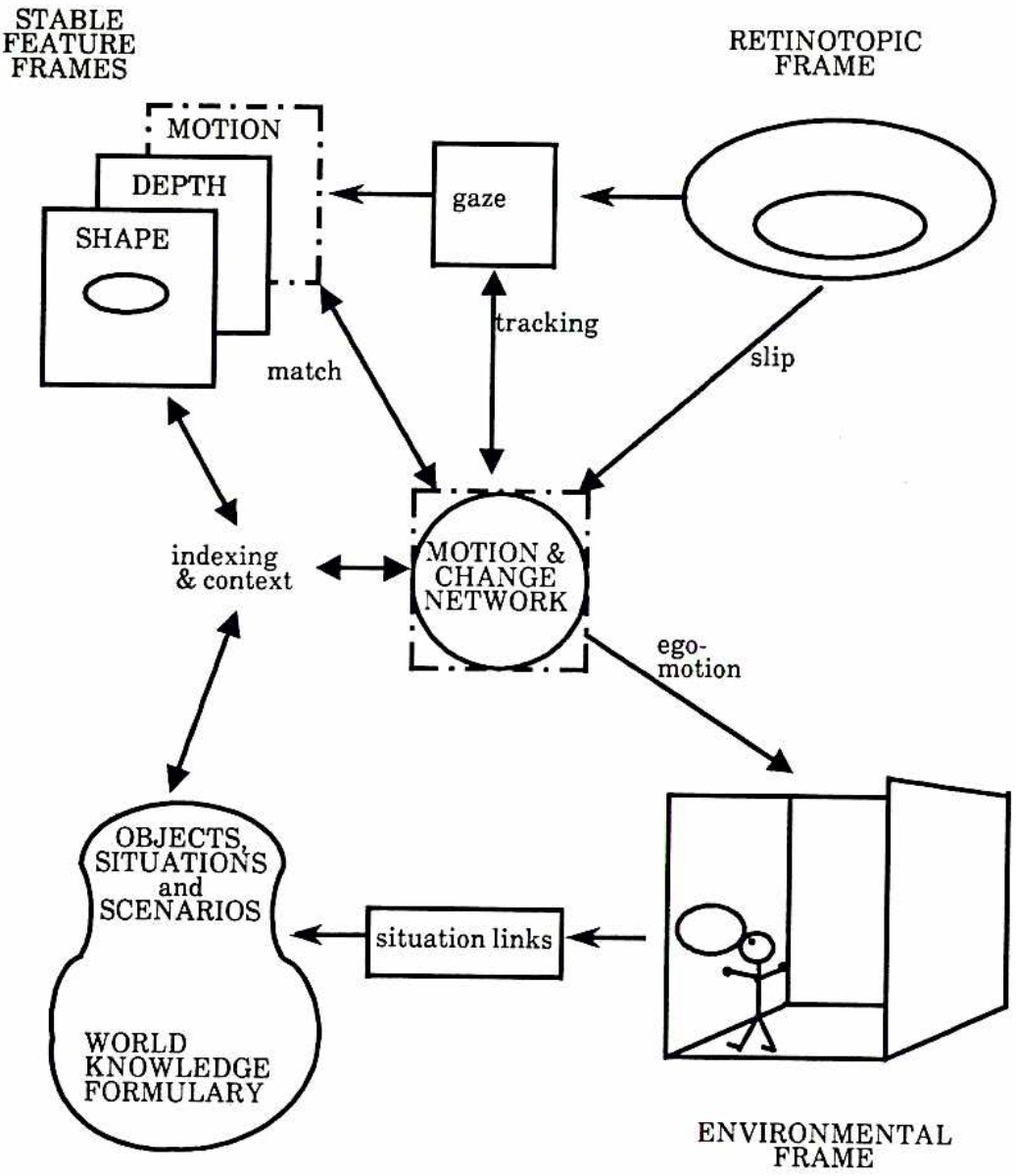


Figure 9: Four Frames and Temporal Change

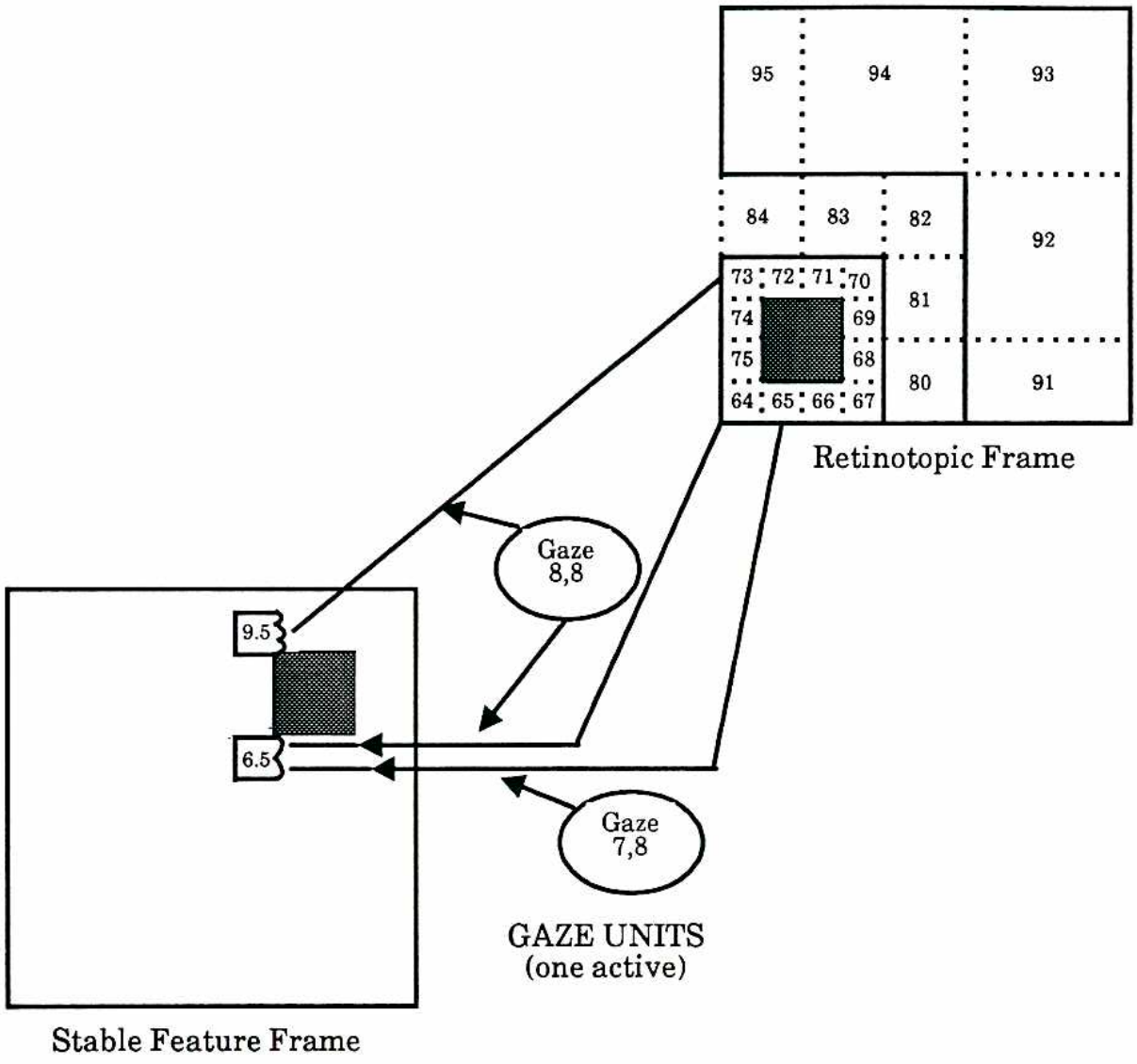

95
94
93
84
83
82
92
73
72
71
70
81
74
69
75
68
64
65
66
67
80
91
Retinotopic Frame
Gaze
8,8
9.5
6.5
Gaze
7,8
GAZE UNITS
(one active)
Stable Feature Frame

Figure 10

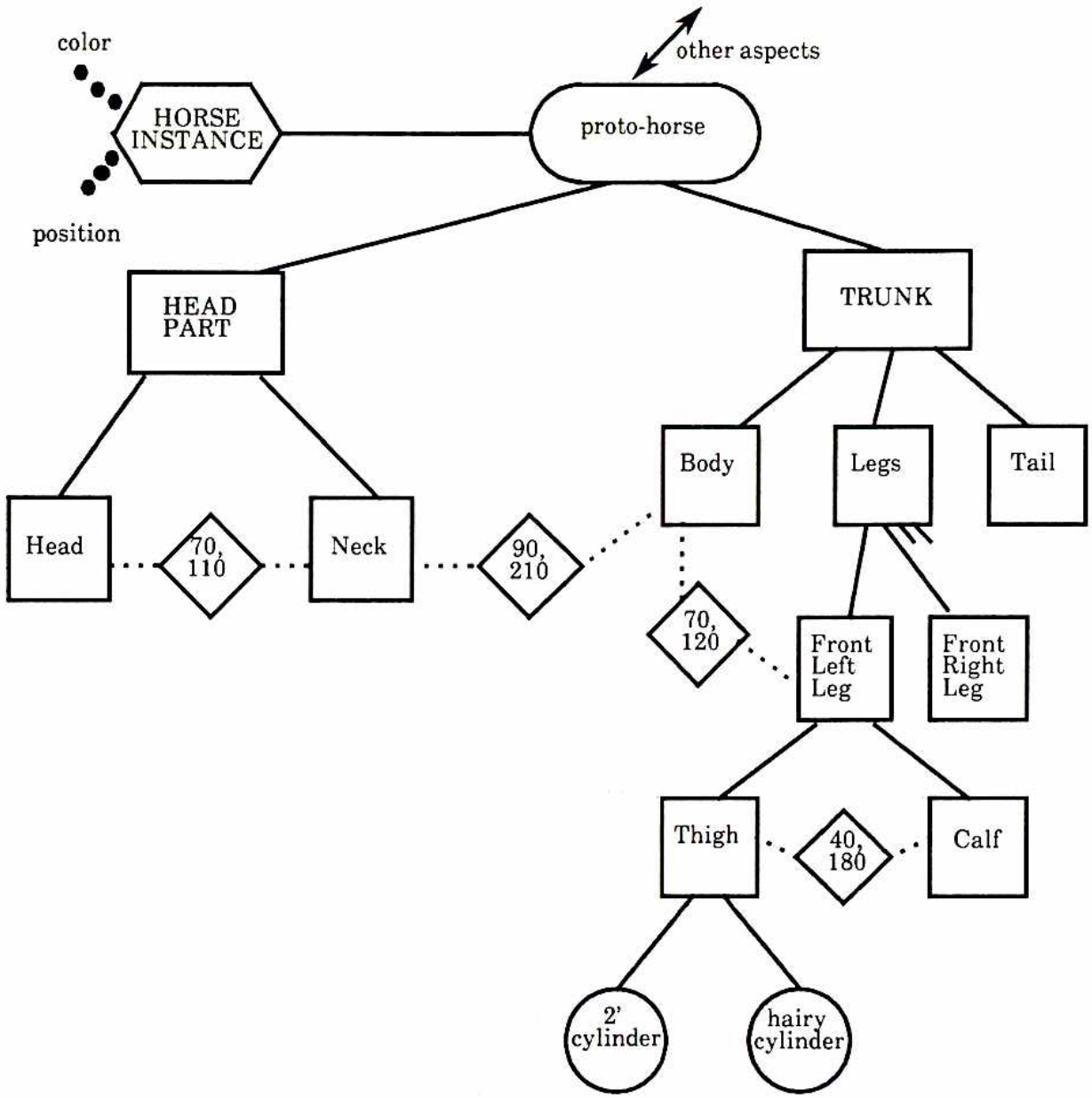

color
other aspects
HORSE INSTANCE
proto-horse
position
HEAD PART
TRUNK
Body
Legs
Tail
Head
70, 110
Neck
90, 210
70, 120
Front Left Leg
Front Right Leg
Thigh
40, 180
Calf
2' cylinder
hairy cylinder

Figure 11

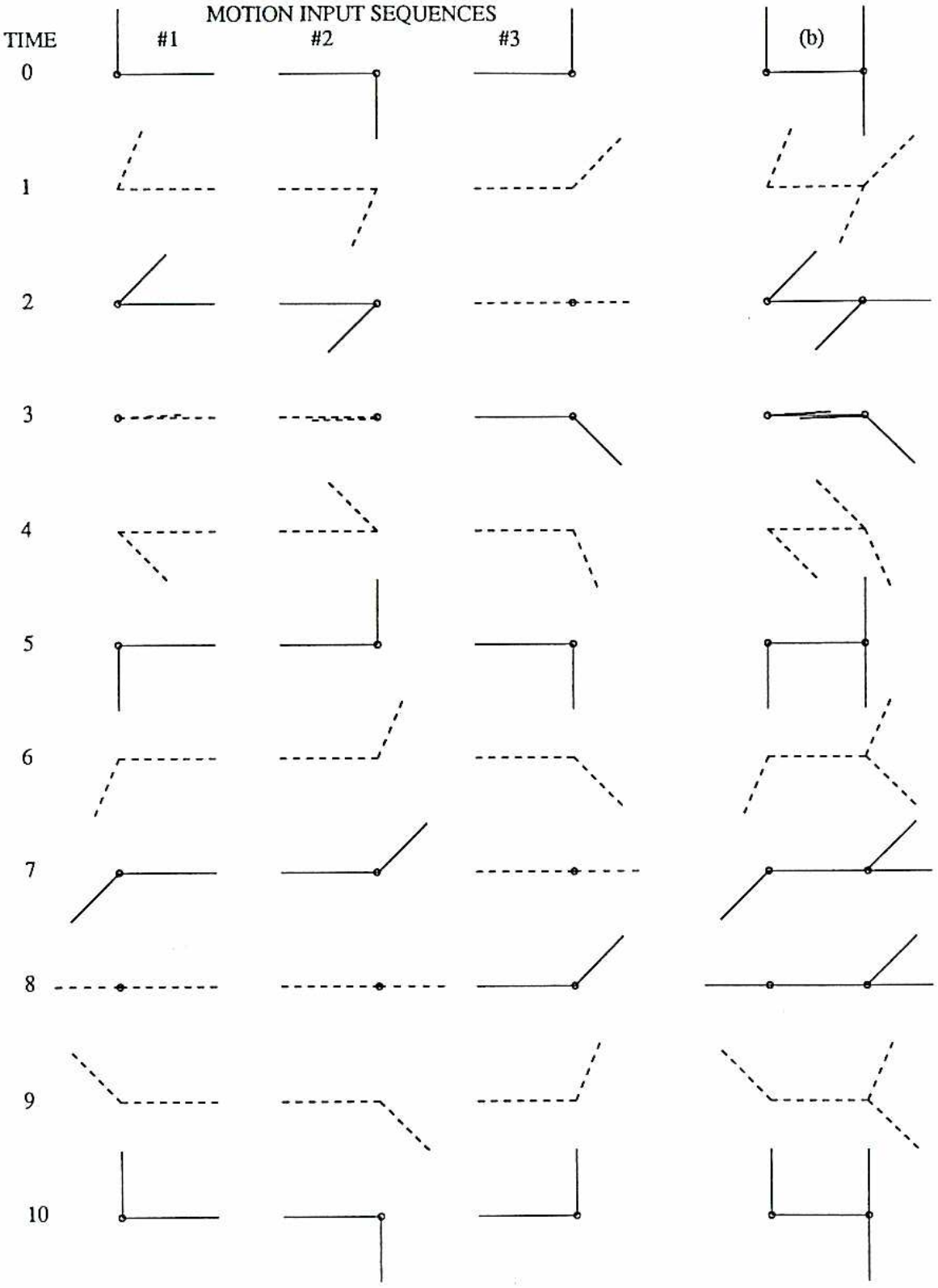


Figure 12

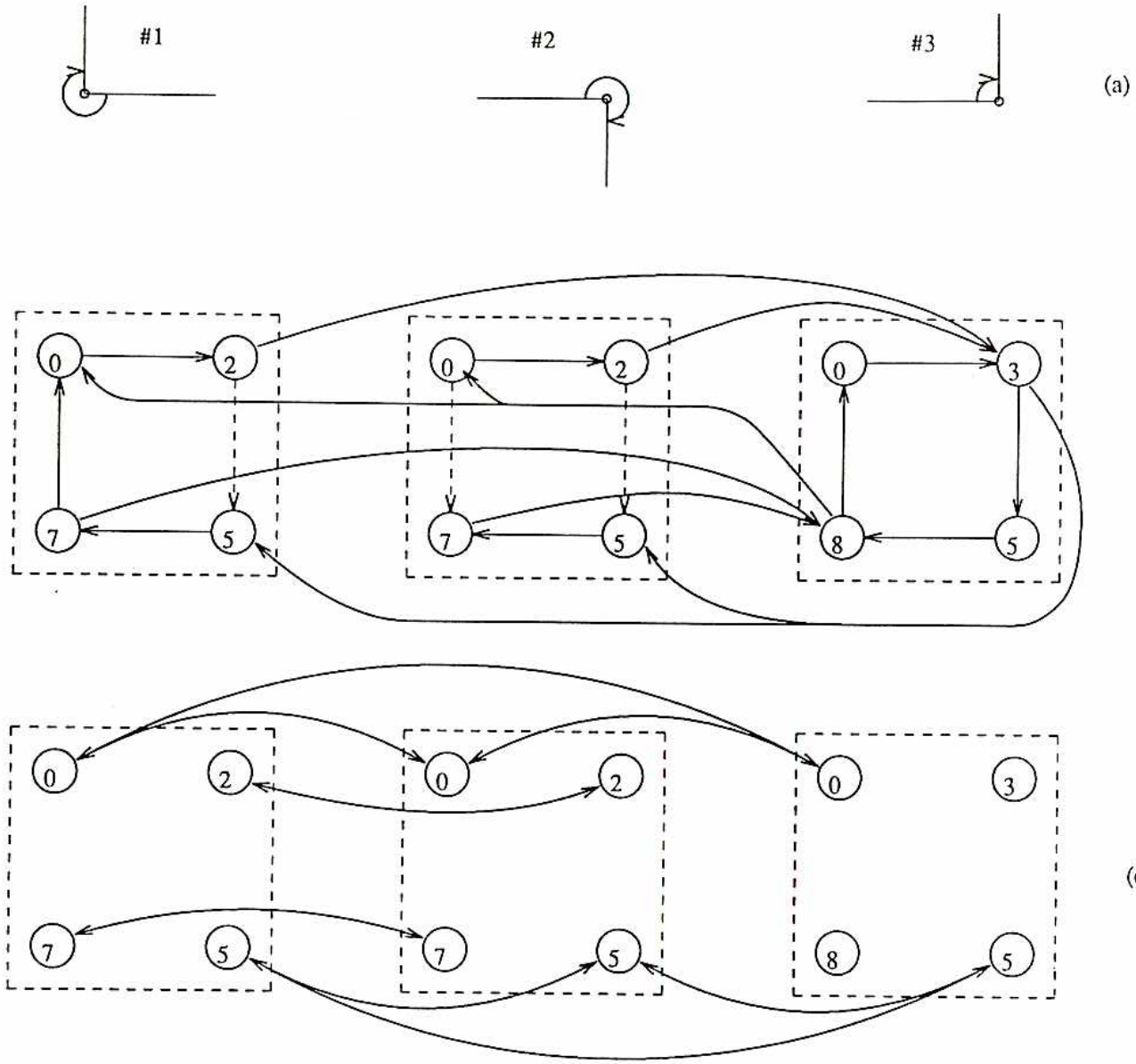


Figure 13

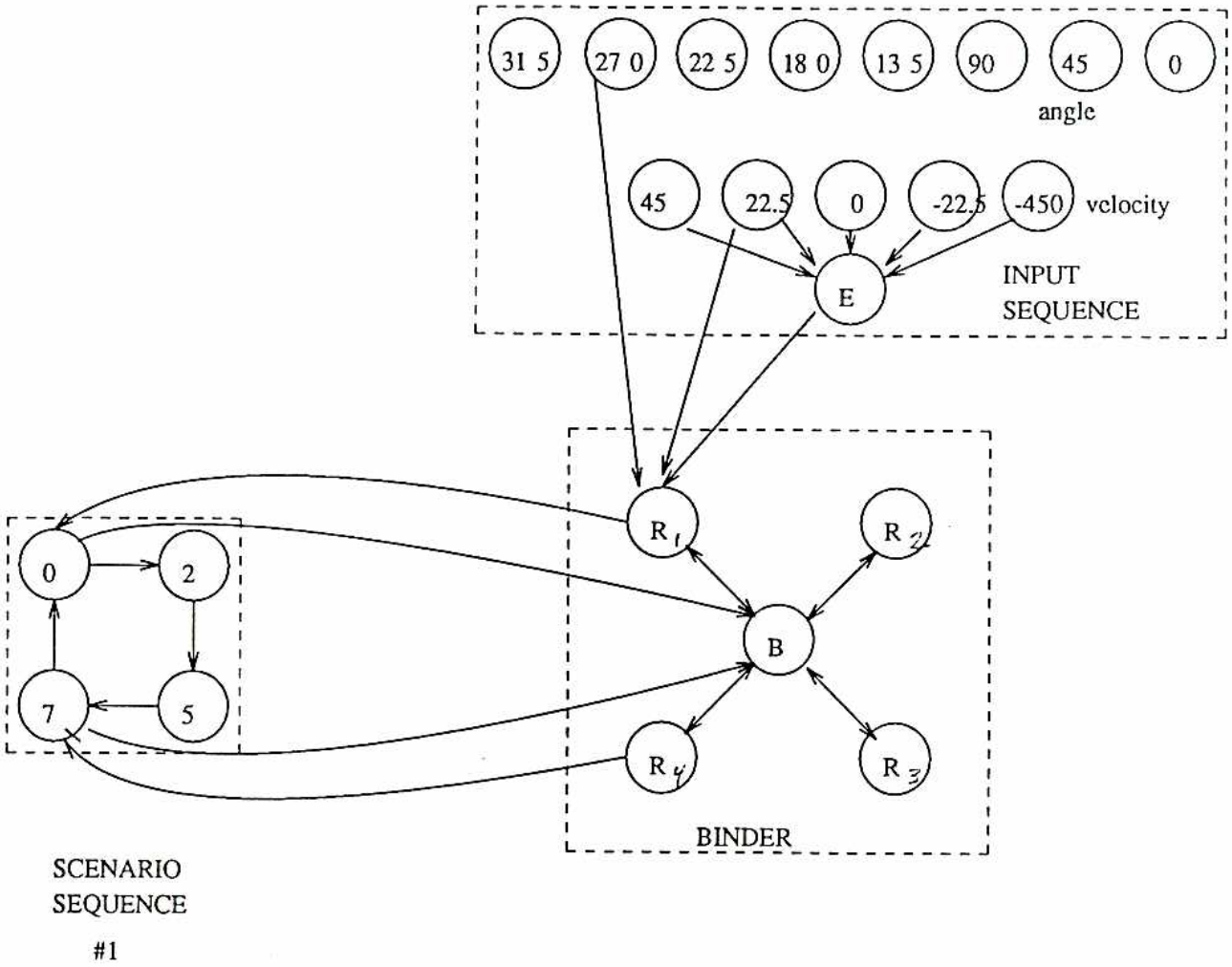

31 5
27 0
22 5
18 0
13 5
90
45
0
angle
45
22.5
0
-22.5
-450
velocity
E
INPUT
SEQUENCE
R1
R2
B
R4
R3
BINDER
0
2
7
5
SCENARIO
SEQUENCE
#1

Figure 14

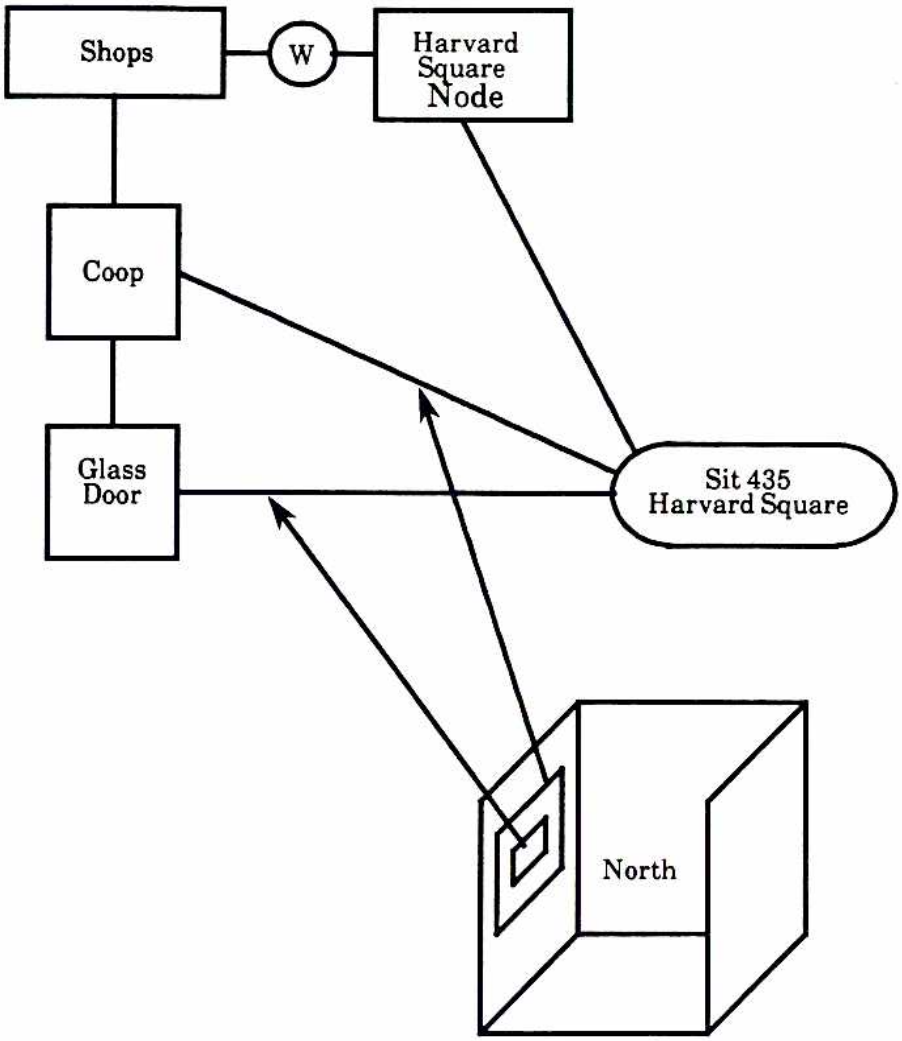


Figure 15: Two Environmental Frame units of different scales activate different objects in Sit 435 = Harvard Square