Computer Science Department Technical Report
Machine Perception Laboratory
University of California
Los Angeles, CA 90024-1596

AN ARCHITECTURE FOR VISUAL DIRECTION CONSTANCY: TOWARDS SHAPE FROM MOTION

Michael Stiber
Josef Skrzypek

April 1990
CSD-900017

MPL

Machine
Perception
Laboratory

UCLA
Computer Science
Department

An Architecture for Visual Direction Constancy: Towards Shape From Motion

Michael Stiber Josef Skrzypek

MPL-TR 90-1

MICRO-Hughes Grant #541122-57442
ARO Grant #444251-25519

MPL

Machine
Perception
Laboratory

top-down

bottom-up

An Architecture for Visual Direction Constancy: Towards Shape From Motion

Michael Stiber   Josef Skrzypek

Machine Perception Laboratory *
Computer Science Department
University of California
Los Angeles, California 90024

Abstract

Projections of a moving target on the retina undergo continuous shifts, especially during egomotion. Without minimal compensation, such as retinal to egocentric conversion, it would be difficult to realize a meaningful interpretation of the environment. We present a model that can perform such a conversion, without resorting to learning schemes, and we argue that its structure is consistent with the known anatomy of the visual system. The model preattentively discounts changes in visual direction, allowing attentive processes to generate a stable and consistent internal representation of the surrounding world. Simulation results from the structured model suggest that the transform demands little computational overhead and is therefore easily reconcilable with visual processing areas that may be performing other operations, such as motion analysis.

*We sincerely acknowledge support by MICRO-Hughes Research Laboratories grant #541122-57442 and ARO grant #444251-25519.

1 Introduction

One of the principal aspects of vision is perception of relative changes in the light stimulation pattern striking retinas. It is from these changes, caused both by movement of objects and our own motions, that we determine overall object shape and its location in three-dimensional space. This presents a problem, namely, how does one build a consistent, stable internal representation [1] of one's environment when the input stimuli are changing continuously, as a result of exploration of the world? Additionally, how does one decide which stimuli merit special exploratory attention, and what sort of mechanism is used to direct this attention? Information related to motion can be used in different computational tasks often indirectly related to the perception of the moving object itself [2]. This motivated many previous studies involving the *oculomotor system* [3,4,5], *direction constancy* [6,7], *position constancy* [8,9,7], and *attention* [10,11,12,13,14,15].

Our ability to build an internal model of the world allows us to explore it [16] — necessary because perception is not just passive interpretation of sensory data, but active exploration of the environment for the purpose of extracting information necessary for survival [17,18]. Exploration, as opposed to passive interpretation, is the only way to operate in unconstrained environments where one cannot enumerate *a priori* all future situations [19,20,21]. This capability is central to "general-purpose vision", which is a long-term goal of computer vision research [22].

From an evolutionary point of view, the problem of synthesizing internal models can be related to the complexities of organisms' behavioral repertoires [23]. A primitive, stationary organism senses any spatio-temporal change in stimuli as object motion, and the location of a point in sensor array coordinates uniquely specifies the stimulus location in environmental coordinates.

More complex behavior requires increased visual acuity, which in turn implies retinal specializations, such as a fovea and nonuniformity of the sensor array. Hence the need to reposition the sensor array to gaze in different directions. Consequently, position on the retina no longer uniquely specifies environmental coordinates and the position of the eye must be accounted for [24,25,26]. This requires a conversion of sensory stimuli from retinocentric to egocentric coordinates.

A fully mobile animal needs to account for motion of its entire body, in addition to sensor array motion. This requires the conversion of retinocentric coordinates into full environmental coordinates, in which objects' position in space is one of the "constancies" to be extracted from visual stimuli [27]. The increased flexibility in combining exploration, manipulation, and introspection of the environment implies the need for a general attention mechanism capable of guiding motor and mental operations in environmental space, [28].

Consequently, we may view visual computation within one or more of three different coordinate systems — retinal, egocentric, and environmental. In this

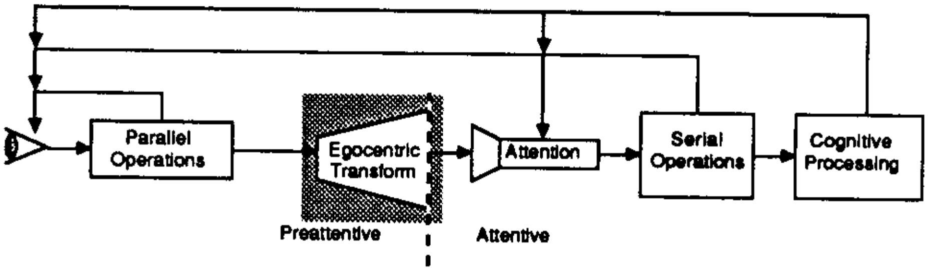


Figure 1: A block diagram of a model visual system, illustrating preattentive/attentive, parallel/serial, and afferent/efferent dichotomies. The egocentric transform is the focus of this paper.

paper, we focus on the transformation of visual information from retinal to egocentric coordinates. Our model is based on connectionist principles, however, the mechanisms are consistent with current anatomical and physiological knowledge of biological vision systems.

The functional subdivisions of our model are summarized in Figure 1. This block diagram contains modules that compute visual direction and position constancies. It is composed of preattentive processes that operate in parallel on the entire retinal image and attentive processes that compute detailed shape information based on some small part of the sensory stimulation. As a model of a general-purpose vision system, which actively explores its environment, it can also be divided into afferent systems that sift through sensory data, and the efferent output of attentive processes which select new data.

2 Visual Coordinates in Retina and Parietal Cortex

The retinal coordinate system is defined by the arrangement of photoreceptors and the geometry and mechanics of the eyeball. Since all retinal cells, including ganglion cells, are fixed with respect to the photoreceptors, computation within the retina and the retinal output are in the same retinal coordinate system. In contrast, it has been suggested that cells in parietal cortex process input within an egocentric coordinate system. In such a coordinate system, points in the environment are specified by their location relative to the individual's body, rather than relative to their projections upon the retina. In the macaque monkey, cells in area 7a respond to both retinal stimulation (which may be the motion of a target on the retina) and eye position, and it is theorized that such a combination may be used to transform retinal coordinates to egocentric coordinates [29,30]. The receptive fields of these cells were mapped out while the animal was fixating in a constant direction [29]. By systematically changing the fixation point, and measuring the cells' response to stimuli at the same retinal location, the authors found that both retinal and eye position information could be modeled as a linear combination.

Furthermore, using the back-propagation algorithm, Zipser and Andersen [30] were able to "teach" a three-layer artificial neural network to simulate a conversion of target retinal position and eye position into target spatial position. After training, they found that the receptive fields of "hidden" units of the network resembled those of 7a neurons, suggesting their role as an intermediate stage in a conversion of visual information from retinal to spatial coordinates.

We developed a model which also translates information from retinal to egocentric coordinates, without resorting to "learning" or error minimization schemes. The functionality of our model arises from the structure of interconnections between model neurons and/or between various known visual centers.

Feldman [1] suggested that four different computational domains were sufficient for the synthesis of stable visual perception. In the context of this proposal, our model implements the transformation from Feldman's "retinotopic frame" to his "stable feature frame", with some important differences. Feldman's retinotopic to feature frame mapping required complete, random interconnections from each retinotopic neuron to every feature neuron to which it could possibly map, despite neuroanatomical evidence that suggests cortical interconnections to be highly specific [31]. Furthermore, it appears that parietal cortex is more likely to be the location of a "stable feature frame" [32] than extrastriate cortex (as Feldman had suggested). In contrast, our model is composed of neurally plausible structures with highly specific interconnection patterns. The number of these interconnections is nowhere near that required by complete interconnection between retinocentric to egocentric neurons. This is a result of a hierarchical organization, which performs the retinocentric to egocentric transform in stages.

2.1 Anatomical Correlates of Model Function

Several of the vision-related areas of the brain, along with their major interconnections, are schematized in Figure 2, adapted in part from [33,2]. We group these areas into four rough categories (as indicated by different shading): I low-level motion analysis, II figure/ground segmentation, III detailed three-dimensional shape analysis, and IV internal representation construction. The flow of visual information between these areas has been shown to split into two major pathways: one projects to temporal cortex and is concerned with detailed analysis of objects, and the other proceeds to parietal cortex, dealing with space perception and the guidance of operations in space [34,2,35,36,37].

It seems reasonable to assume that the location of the egocentric transform would be in the path dedicated to spatial operations. However, the anatomical localization of the egocentric transform could depend on whether the conversion is computed in a distributed fashion or whether it is accomplished within one specific visual center. The transform could be distributed across several preattentive visual areas, such that each area represents one stage of the overall transform, in addition to performing other operations, such as optic flow, color, etc. It is also possible that there is only a single area which performs the entire conversion — an area dedicated to direction constancy. While at present there is not enough evidence to eliminate either possibility, we can narrow the focus to include visual cortical areas V3, MT, 7a, and MST [38,28,29], as shown in Figure 2 by the shaded areas.

The results of a complete egocentric transform would then allow attentive processes to direct attentive motor operations toward targets in spatial coordinates. This notion is supported by recent findings which suggest that targets are specified to thalamic nuclei in spatial coordinates [4].

3 A Model of the Egocentric Transform

The objective of our model is to simulate the retinal to egocentric coordinate conversion. To simplify simulations we ignore the many types of wide-area preattentive visual processes that must be operating concurrently. It is assumed that operations to compute preattentive color, texture, motion, etc. are being performed within the complete visual system, of which our model is just one part. The model is composed of two-dimensional layers of "neurons", each of which receives input from a linear arrangement of neurons in a previous layer. Inhibitory inputs are taken as all-or-nothing, that is, inhibition on one "dendrite" blocks any excitatory input from that dendrite. Therefore, a particular visual input can be either connected or disconnected.

To further simplify simulations, we use discrete frames of visual information — however, the model's behavior would be equivalent if it were stimulated by the "smooth" motion inherent in the real world.

We have been motivated in our efforts by the following observations from neuroscience. The receptive fields of visual neurons increase in size as we move towards more central visual areas [39]. This suggests that parietal neurons can potentially receive input from a very large number of photoreceptors. There is evidence which suggests that parietal neurons can integrate eye position and visual information [29]. Finally, receptive fields of visual neurons can have their properties modulated by inputs from other visual areas [39].

3.1 System Architecture

Changes in the retinal location of visual stimulus can be caused by either object motion or by eye motion. In order to reconstruct the egocentric locations of objects, one must differentiate between these two possibilities. The transformation must be information preserving; projections of moving objects on the retina

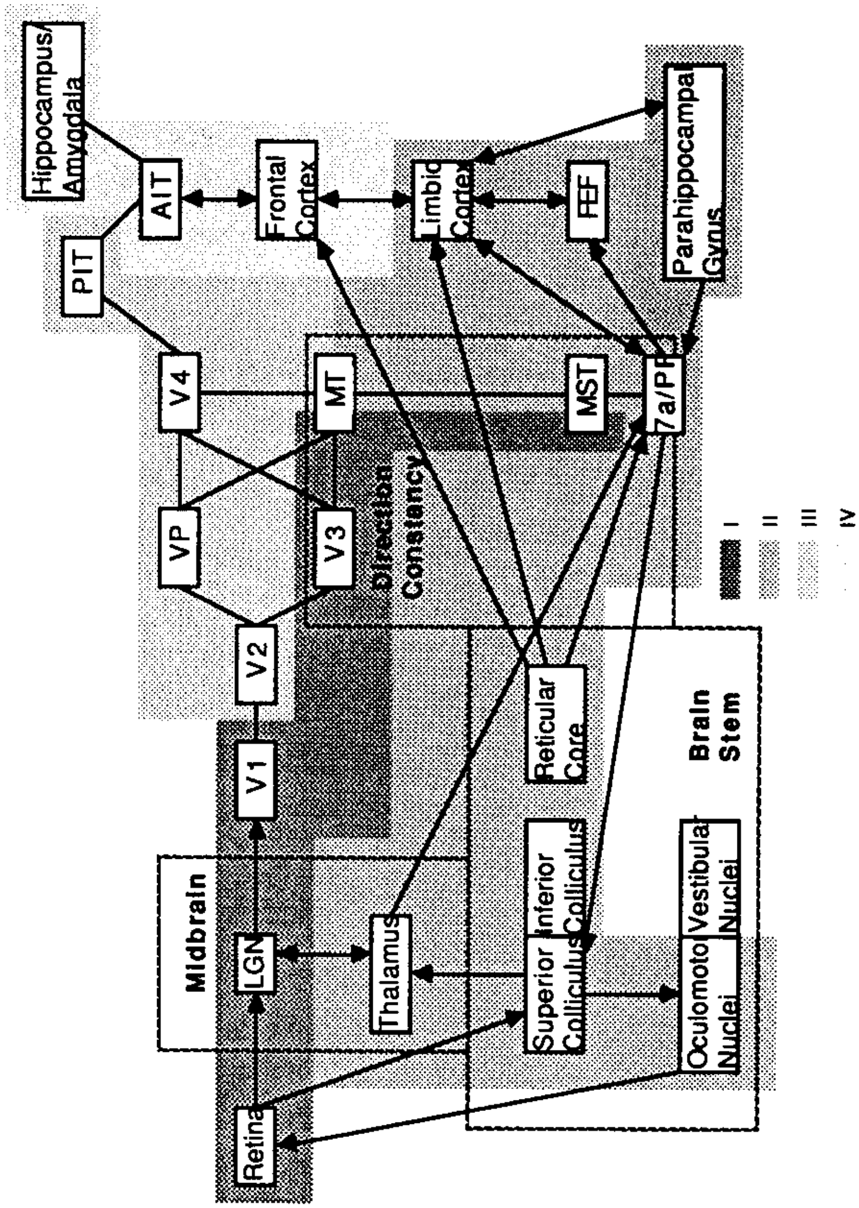


Figure 2: A greatly simplified block diagram of the visual system shows major areas in retina, brain stem, midbrain, and cortex, and their relations to motion and direction constancy computation. I. low-level motion. II. figure/ground segmentation. III. detailed shape. IV. environmental modeling.

must produce moving patterns of activity in the transformed space, in proper spatial relationship to each other. The transformation must discount artificial changes in the retinal location of projections of stationary objects.

Figure 1 shows how this egocentric transform fits into the context of biological vision. The input to the transform (from the retino-geniculo-cortical pathway) is the result of processing images by neurons with symmetrical receptive fields with center/surround antagonism — typical of the responses of the cells in the retina, LGN, and striate cortical areas (for a review, see [35]). The output of the transform could correspond to either of Treisman's two hypotheses of attentional mechanisms [40,10,41]. In the first hypothesis, feature extraction modules send their outputs to a "master map of locations", which indicates only where features are, but not what they are. In the second hypothesis, features are initially combined in the master map, and are later separated apart. The difference is that, for the latter, preattentive conjunctions of features can be made (since the features are initially together in the master map). For the former, features are separated apart before reaching the master map (and, by extension, attentive processing). Conjuction of features would then require attentive processing.

The output of our model corresponds to the master map of locations in the former, and the feature module outputs in the latter. The crucial point is that our model suggests that either the master map of locations is in egocentric coordinates (in the first case), or that the feature module outputs are in egocentric coordinates (in the second case).

3.1.1 Principles of Operation

Figure 3 shows a functional block diagram of our model, examining a stationary scene. At time t_1, the camera fixates on one object. Visual information that is processed preattentively undergoes a retinal to egocentric transform, which places the features extracted into their correct locations in a world map. Attention, acting upon information within this map (and other, nonvisual information), decides to shift camera direction to fill in the feature maps with more information. In the new direction, preattentive feature extraction processes operate just as they always have — those computations are based on retinal information, and do not take into account shifts of point of view. However, the egocentric transform assures the system that the second object's contribution to the description of the world is kept distinct from the first object's.

A block diagram of the architecture that accomplishes the coordinate transform is shown in Figure 4. It is composed of two types of processing layers — shift layers and control layers. Shift layers process images from the sensor array, while control layers process sensor array position information. The sensor array/shift layer pathway corresponds to the geniculocortical system for processing visual information [35,33,37]. The control layers correspond to parts of the brain stem – superior colliculus – thalamus – cortical system for eye position

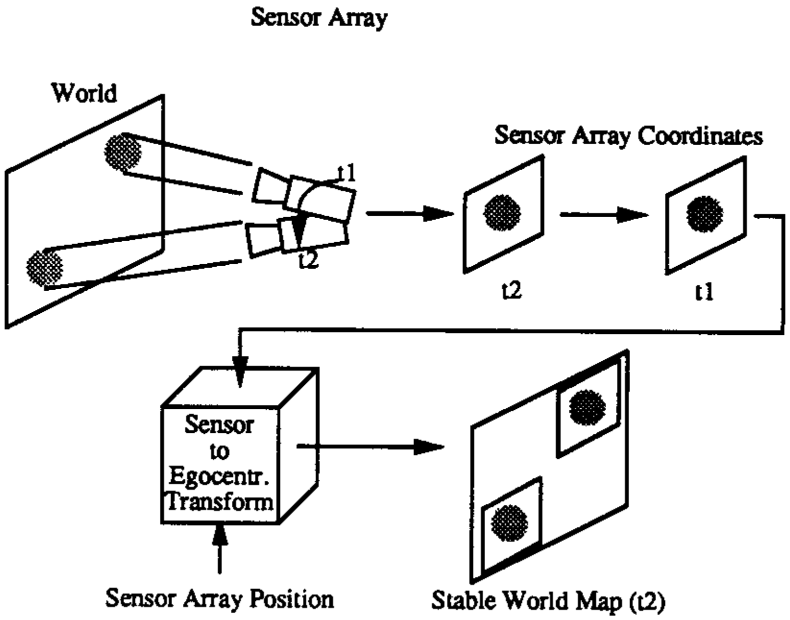


Figure 3: A "cartoon" example of the egocentric transform. A moving camera takes two nearly identical images in succession, which, when placed within egocentric coordinates, are clearly of two different objects.

data [4,42].

Shift layers are arranged to form a truncated pyramid. Each layer can contain an image at some stage of the transform. The bottom layer is essentially in sensor array coordinates, perhaps receiving its input from LGN, and more likely from the early visual cortical areas, such as V1 and V2. The topmost layer is in egocentric coordinates, providing its output to areas in parietal cortex that coordinate operation of the visual system with locations of objects. Intermediate layers perform partial shifts of the images in a single direction at any one time, with higher layers performing larger magnitudes of shifting. The composite shift of the shift layers serves to place images in their proper location in the egocentric map (see also Feldman's "stable feature frames" [1]). Since the sensor array can be oriented in almost any direction, it must be possible to route the output of a processing element in the bottommost layer to just about any PE in the topmost layer. The shift layers accomplish this mapping operation.

Control input to shift layers is equivalent to the eye position information transmitted to superior colliculus and the thalamic internal medullary laminar complex [4,42], and presumably on to the parietal cortex via heavy thalamoparietal interconnections [43]. Control layers determine the direction and magnitude of image shift for each shift layer. There is one control layer for each shift layer. Sensor array position is provided to the topmost layer, which determines the shift direction and amount for the topmost shift layer, which is the largest magnitude shift. Successive control layers break the position into

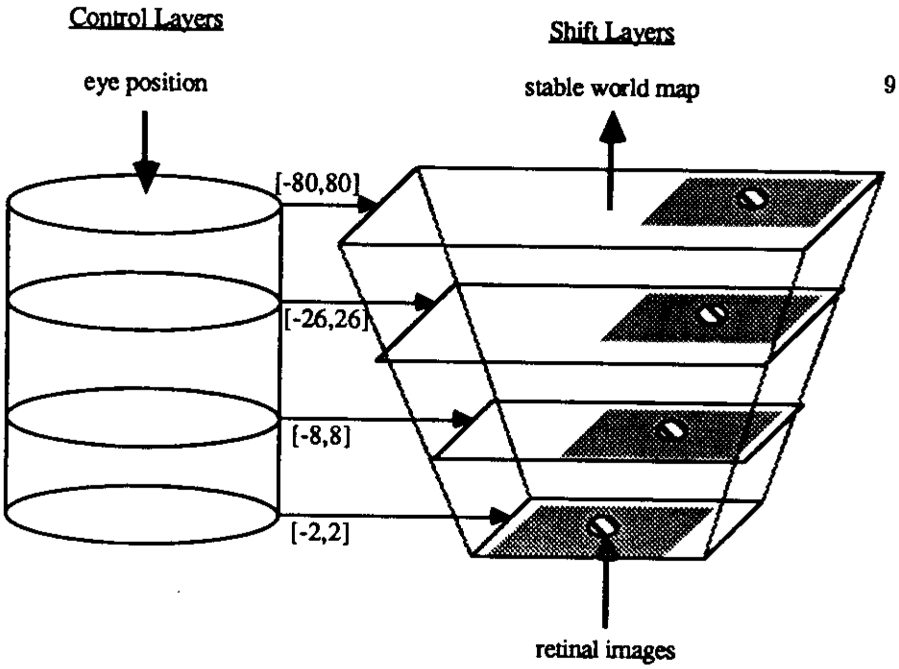


Figure 4: A block diagram of a neural network architecture for performing a retinal to egocentric coordinate conversion. *Control layers* use eye position to compute image shifts. *Shift layers* perform successive shifts of a retinal image, resulting in its placement into the appropriate location in the egocentric map.

smaller magnitude components, corresponding to the smaller magnitude shifts performed by the corresponding shift layers. Thus, the current sensor position is decomposed into components which correspond to shifts performed by shift layers. The net result is to place an image into the location in the egocentric map dictated by the sensor array orientation.

3.1.2 Shift Layers

A series of shift layers are used to translate the sensor array image to its proper location in egocentric coordinates. Although this could be simulated as a one-step process, it would require each PE in the egocentric map to be connected to almost all of the PEs in the sensor array map. This would be difficult to realize in a hardware implementation. Additionally, the shift elements are rather simplistic — many additional functions, such as sub-pixel resolution, could be performed by them.

A shift layer can perform a unidirectional shift of an entire image. Since our architectural paradigm only allows connections over some limited local area, arbitrary shifts by one layer are not possible. That is why a series of layers are used, and why successive layers are interconnected so that they perform increasing magnitudes of shifts. This is comparable to increasing the size of receptive fields — the increasing shift magnitude is produced by interconnection between layers and the multiplicative effect of successive shifts. For example, if every PE is connected so that it has five inputs, and can perform shifts of $\{-2,-1,0,1,2\}$ pixels, the ensemble of shift layers are capable of the following shifts:

- The bottommost layer shifts in increments of 1, within a range of -2 to +2.
- Layer 2 shifts in increments of 3, with a composite range of -8 to +8.
- Layer 3 shifts in increments of 9, with a composite range of -26 to +26.
- Layer 4 shifts in increments of 27, with a composite range of -80 to +80.

Subsequent layers follow this pattern of increasing magnitude, corresponding to powers of 3 in this example. Except for their increasing size, the shift layers are all identical.

As a more realistic example, consider the human visual system. At the retina, the field of view is approximately 200° [44], and the resolution (at the location of the simple cells in the visual cortex) ranges from 0.1° at the fovea to 3° in the periphery [45]. If we take the maximum resolution and field of view to arrive at an upper bound on the size of visual maps, that would be equivalent to a 2000 photoreceptor by 2000 photoreceptor sensor array. Let us assume that the egocentric transform is also used to increase maximum acuity through an interpolation process. In humans, maximum visual acuity (not hyperacuity) is approximately 0.01° [46]. Though eye motion may not cover all space around an individual, we assume the egocentric map to be 360° in extent to get an upper bound on map size. Additionally, we will assume that all images are mapped onto square visual areas, resulting in an upper bound on the size of the egocentric map of 36,000 cells by 36,000 cells $(\frac{360 \text{ degrees/side}}{0.1 \text{ degrees/cell}})$. Of course, visual acuity falls off as distance from the fovea increases. Additionally, a 36K x 36K square has greater area than a spherical 360° egocentric map. For these reasons, the figures of a 36K x 36K square map will serve as upper limits on this architecture.

The egocentric transform must then be able to shift a 2K x 2K input image anywhere within a 36K x 36K egocentric space. It has been suggested that, in macaque, cells in V1 which receive magnocellular afferents have receptive fields which cover an area corresponding to 18 receptive fields at the LGN [47]. The parvocellular afferent recipients in V1 are arranged in submodules that exhibit high cytochrome oxidase activity, and that these submodules could receive between 10-15 afferents each [47]. Let's assume that each cell in a shift layer is connected to approximately 10 cells at the previous layer. The first shift layer would then shift its input one cell at a time. The second layer would perform shifts in increments of tens of cells, and the third by hundreds. The fourth and final layer would be capable of effecting composite shifts on the order of thousands of cells.

Thus, the egocentric transform can realistically be accomplished by four shift layers, with cells in each layer connected to a small number of cells in a previous layer. The total number of cells required would be dominated by the number of cells used in the final egocentric map. It is not unreasonable, then, to assume that shift operations performed by each shift layer actually correspond to separate visual cortical areas. Each area would perform a stage of the computation of preattentive feature maps, as well as a stage of the transformation of the map information from retinocentric to egocentric coordinates. In that case, our model would argue in favor of Treisman's "late master map of locations" hypothesis, in which the feature maps feed their outputs to a master map (corresponding to our egocentric map).

A "lumped" model of the early geniculocortical system was employed for simulation convenience. While the sampling of visual space by the human retina could be modeled as a roughly 2000 by 2000 photoreceptor image, we have used 64 by 60 pixel input images, so that one pixel (or shift element) corresponds to 1024 photoreceptors. An additional simplifying assumption is that our input "retinal" images are of uniform resolution.

Each processing element within the shift layers is individually capable of shifts in only one direction, as illustrated in Figure 5. Each shift element receives a linear arrangement of image inputs, any one of which can be selected to be that element's output. Thus, the shift element pictured in Figure 5 is capable of shifts: $\{-2, -1, 0, 1, 2\} * \mathit{magnitude}$, as described above. The actual input selection is performed by the control layer signals. Active signals from the corresponding control layer "shuts off" shift layer inputs. By inactivating one control to a layer's shift elements, one image input is selected.

To accomplish shifts in arbitrary directions with elements capable of unidirectional shifts only, multiple shift elements per location are need. Each of these elements is capable of shifts in a slightly different direction (8 directions are used in the simulations). By using a sequence of shift element of appropriate orientations, shifts in arbitrary directions can be accommodated. The bottom of Figure 5 shows the multiple shift elements at one location in a shift layer. Selection of the appropriate shift direction and magnitude is done by the control layers.

The shift elements are arranged in a rectangular array. At each location in the array, there is a set of shift elements, one for each direction. All of the shift layers are identical in construction (except for size) and interconnection (except for the first layer). Later shift layers are closer to the egocentric map, and each succeeding one must therefore be closer in size to the egocentric map (ie., larger). The final shift layer corresponds in size to the egocentric map.

The elements in a succeeding layer do not receive input from adjacent elements in a preceding layer. For example, for shift magnitudes that increase in steps of 3, an element in layer n takes inputs from every third element in layer $n-1$. Its neighboring elements do the same, but with inputs shifted by

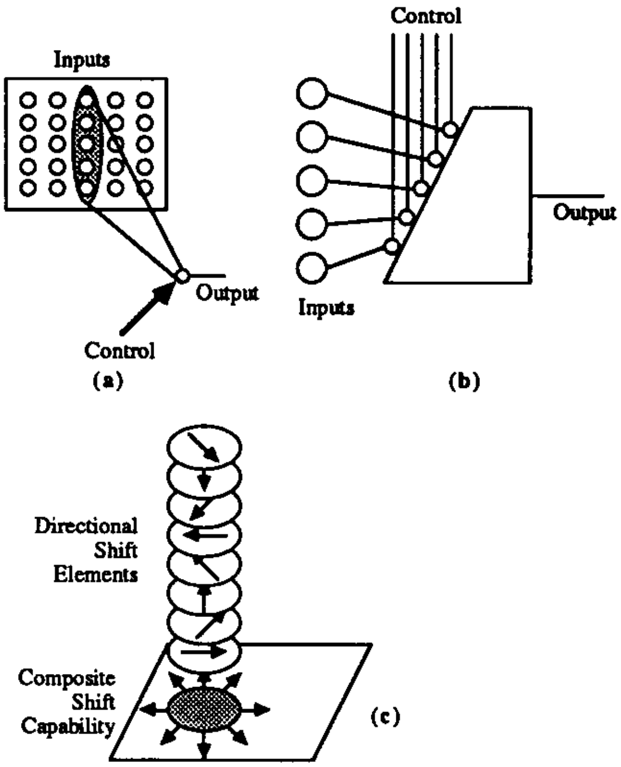


Figure 5: Shift elements receive a linear arrangement of visual inputs from the previous layer (a). Control inputs selectively inhibit individual visual inputs (b), so that the output of a shift element is equivalent to one of the visual inputs, resulting of a shift from -2 to +2, in this figure. At each location in a shift layer, shift elements exist for 8 different shift directions.

1 with respect to the former element. The exception to the scheme is the first layer. Elements in this layer receive inputs from adjacent sensor array outputs, to effect shifts of magnitude 1.

3.1.3 Control Layers

The function of control layers has in effect already been defined. There is one control layer for each shift layer. They receive information about sensor array location, and send outputs to shift layers. These signals select inputs for shift elements such that a shift of a particular magnitude is performed in a particular direction. Successive control layers break the sensor array location into finer increments, corresponding to the smaller shift magnitudes of the lower shift layers. Their function is quite similar to the Granger-Lynch model of olfactory paleocortex [48], in that it is composed of winner-take-all layers (ie, layers in which only one node responds at any time), which collectively construct a hierarchical decomposition of eye position. Unlike that model, however, our control *layers* are separate, rather than patches in one layer, and the structure of layer interconnections is hardwired, rather than learned.

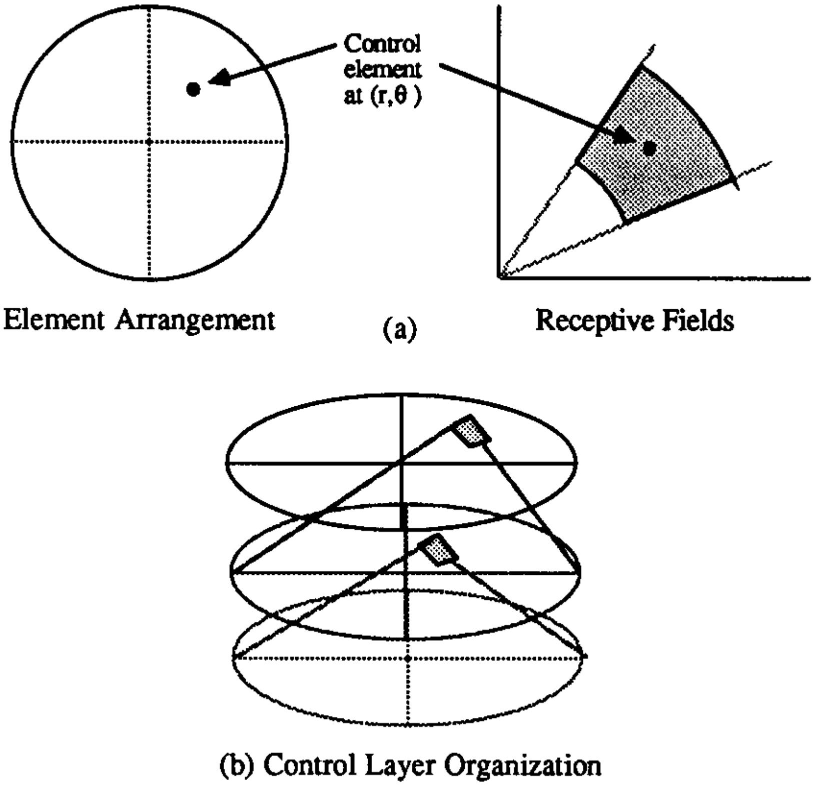
Figure 6 illustrates the organization of the control layers and the receptive field characteristics of individual control nodes. Control processing elements in the topmost layer have receptive fields that correspond to some part of visual space. They are sensitive to position of the sensor array and don't participate directly in processing image data. A control element will produce a low output when the sensor array is pointed in the direction that the element's receptive field covers, as determined by the array position inputs. Thus, when one looks at the entire top control layer, one sees all control elements but one producing high outputs, with the low output corresponding to the region of space towards which the sensor array is pointed. The result is that, for any shift element at that level, all inputs but one are inhibited.

Subsequent control layers break the spatial position of the sensor array into finer divisions, corresponding to the smaller magnitude shifts produced by the shift layers. Thus, a receptive field of one element in a control layer becomes the entire space to which the next control layer is sensitive, as shown in Figure 6. Elements in that layer are then comparing the direction of the sensor array in space with the approximation to which the previous control element corresponds. When a control element produces a low output, that means that the actual sensor array direction falls within that element's receptive field. However, that control element activates a shift by the corresponding shift layer in a particular direction with a particular magnitude. The difference between the sensor array location and the shift performed can be thought of as residual error, to be corrected by subsequent control/shift layer combinations at progressively finer resolutions. Therefore, the next control layer's elements operate within the space represented by the previous receptive field, with the actual shift performed corresponding to the origin.

For example, if the sensor array is pointed up and to the right at an angle of 45° from the horizon and 10° from straight ahead, then a control element corresponding to 45° in each control layer would be inhibited. This means that inputs to shift elements at an angle of 45° to their location would be left uninhibited. The result is a 45° shift in each shift layer. The shift magnitude would be adjusted in each layer to accomplish the 10° cumulative shift.

3.2 Methods and Procedures

The results presented here are all based on simulations at the UCLA Machine Perception Laboratory. These simulations use a neural network simulation environment called SFINX (Structure and Function In Neural conneXions) [49,50], running on Ardent Titans and other UNIX machines. SFINX allows simulation of both "randomly" and regularly interconnected networks, where individual



(b) Control Layer Organization

Figure 6: Control elements convert eye position into the successive shift directions and magnitudes required by the shift layers. Each element has a receptive field which is responsive to a certain range of eye position magnitude and direction (a). control layers provide finer resolution by comparing the coarser approximations of the previous layer with the input eye position (b).

"neurons" can be simulated at almost any level of detail, from a set of difference equations to a simplified weighted-sum model. To accomplish this simulation, functions must be written that simulate individual neurons, with SFINX taking care of applying that function for each neuron in the net. The functions define not only the operation of individual elements, but also the interconnection between elements (in the case of regular interconnections). SFINX includes facilities for displaying the internal state of elements as images, when regular interconnection is used. Additionally, the state may be saved to a file for further processing.

To design simulations which would run in reasonable time, and to make interpretation of results easier, certain simplifications were made to the model. We used a lumped model of the early geniculostriate system, and in some simulations neglected the inhibitory surrounds of the early system, so that results could be viewed as normal images. However, simulations that required measurement of receptive fields did use inputs that had been processed by cells with center/surround receptive fields. Additionally, in our simulation, the egocentric transform does not involve the computation of feature maps. Instead of the output being feature maps, unaltered image intensity information is used.

The proposed vision system is based upon connectionist architectural principles — parallel designs with simple processing elements (PEs) which are connected to other processors within some restricted area (for a review, see [51]). Restricting the area of interprocessor connection keeps the number (and complexity) of such interconnections reasonable, as resources required for interconnections may grow with the square of the interconnection diameter. The processors operate individually on small parts of the problem — it is only as a group that they perform a recognizable function. This is a result of the limited interconnection area, which prevents any processor from having a global view of the task. It also adds an element of fault tolerance to the design, as losing individual PEs results in only small degradations in overall performance.

PEs are arranged into *layers*; two-dimensional arrays of elements. Thus, the data being processed (images) map directly onto a layer of PEs. These layers are "stacked" on top of each other to form complete computing structures, forming a pipeline through which image frames flow — with sequential frames in sequential stages of the pipe. Adjacent layers may be the same size, as in our model's control layers, or different sizes, as in the shift layers.

For the simulations performed here, a regular interconnection scheme is used. In this scheme, two-dimensional buffers are allocated to hold the current state of each PE. A single function is defined for each buffer, which describes the operation and interconnection of PEs within that buffer. For simulations involving vision, a SFINX buffer usually corresponds to a layer in the architecture. Thus, there is a close correspondence between the structure of the simulation and the structure of the architecture, which makes analysis of its behavior straightforward.

The simulation results can be presented in two ways. One is as images, with brightness at each point in the image corresponding to the level of activity of a PE at that location in the layer. The other is as three-dimensional "terrain maps", with terrain height corresponding to PE activity level. Some of the data is captured by adding layers to the simulation whose only purpose is to accumulate the output of individual PEs or entire layers over time. This allows one to see, in one picture, the results of hundreds of test iterations.

The test conditions involve simulation of a static world, where the only motion is sensor array motion. This is accomplished by adding "world" and "sensor array" layers before the first shift layer. The world layer consists of a still image, meant to be taken for the entire visible world (two-dimensional case). The sensor array layer is much smaller. For a particular sensor array direction, that layer takes a rectangular piece of the world image. By changing the sensor array direction, different parts of the world may be "examined" by the system. In the simulations presented here, the world image is 256 pixels across by 240 pixels down, while the sensor array is 64 pixels across by 60 pixels down.

The contents of the sensor array are presented to the lowest shift layer of the architecture — a simulation of the retinal ganglion output to the visual cortex. The sample of the world image taken by the sensor array is passed unchanged, even though the retinal ganglion cells exhibit center-surround receptive fields. This is done to clarify the data collected from the simulation, as the final output is still interpretable as a normal image.

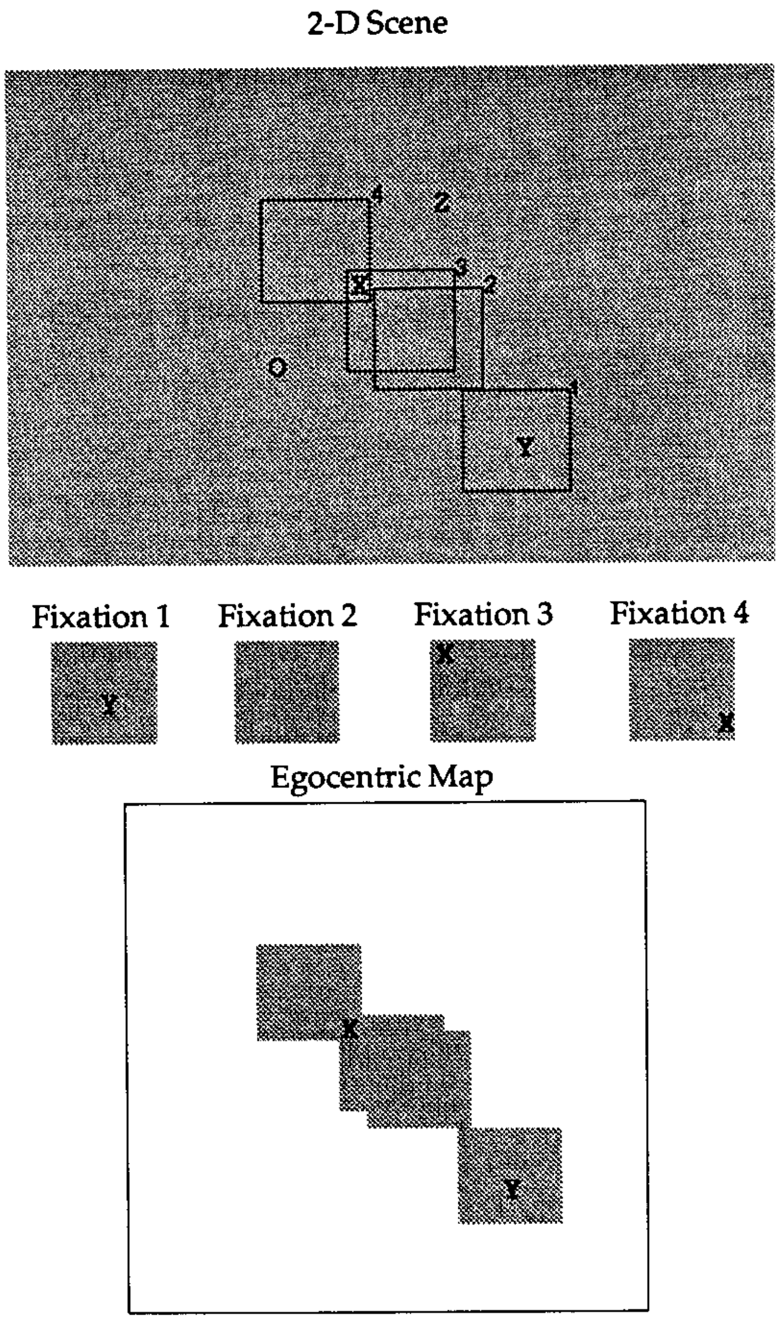
In all simulations, the shift layer simulation code was written to allow for eight different shift directions (simplest with a rectangular pixel arrangement) and any number base for shift magnitude. With the world and sensor array dimensions stated, the most efficient shift magnitude (in terms of memory usage) is five, ie. a range of [-4,+4]. Three shift layers were used, thus the first layer shifts in increments of 1, the second in increments of five, and the third in increments of 25. With a sensor array that is 64 x 60, the first shift layer is 72 x 68, the second is 112 x 107, and the third is 312 x 307. These dimensions accommodate the maximum shifts at each point. All control layers were 5 by 8 (shift magnitudes from 0 to 4; 8 shift angles).

3.3 Experimental Results

Figure 7 illustrates a typical experimental setup. A simple two-dimensional scene is constructed, containing the letters 'O', 'X', 'Y', and 'Z'. The model retina's field of view is a small part of this scene, indicated by boxes drawn upon the scene. Four fixations were made in the test illustrated, corresponding to the boxes numbered 1 through 4. Below the scene image, the part of the scene viewed by the retina in each successive fixation is shown. A final data collection layer was added to the model after the final shift layer, to accumulate shift layer outputs over time. The bottom image in Figure 7 displays the accumulated output of the model for the time period covering the four fixations.

These accumulated outputs show that the fixations are each registered to their proper locations in egocentric space. This results in a reconstruction of those parts of the original scene which were viewed by the retina.

Figure 8 shows a more complex scene. The retina was scanned back and forth across the entire scene, so that all parts of the scene were observed during at least one fixation. By accumulating the shift layer outputs as in the previous experiment, the entire scene can therefore be "reassembled" in egocentric space, as in Figure 9. Careful examination of Figure 9 reveals that the reconstruction is not perfect. This is an artifact of the rectangular arrays used for processing layers, since the diagonal shifts are in increments of $\sqrt{2}$ greater than the horizontal and vertical shifts. A hexagonal buffer system would solve this problem, with a cost of greater simulation complexity and time.



2-D Scene

Fixation 1

Fixation 2

Fixation 3

Fixation 4

Egocentric Map

Figure 7: Result of model operation for 4 fixations of a simple scene. The top picture is a simple scene, with numbered fixation areas. The contents of the retina for the fixations are shown in the middle, and the final egocentric map contents are on the bottom.

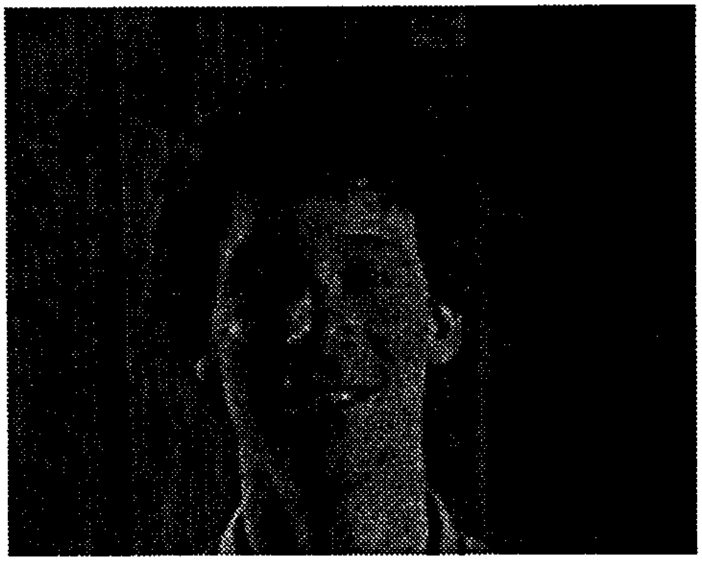


Figure 8: An example image, representing a complex scene.

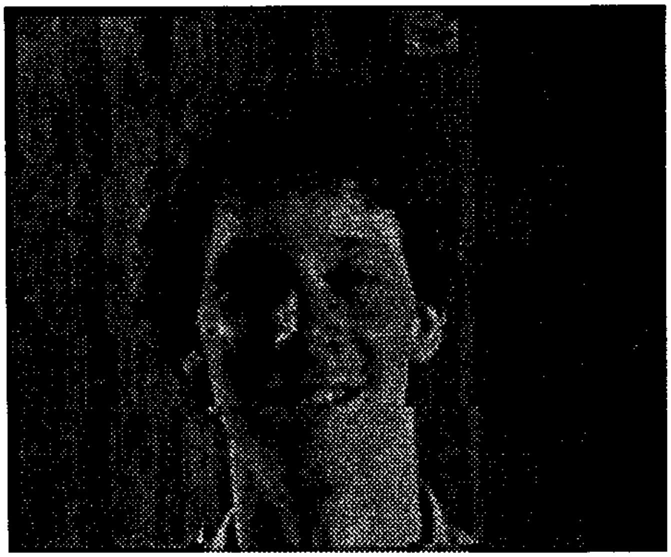


Figure 9: The image in Figure 8 was viewed by a small "retina", which was scanned back and forth across the image. The outputs of the final shift layer were collected to form this reconstruction of the egocentric map.

4 Conclusions

Modeling involves choices: levels of abstraction, generalization, correspondence with observations, representation of observations, etc [52,53]. We have chosen to focus our efforts on building models of biological neural *systems*, rather than models of experimental *data* [54]. Our model details a computational structure that can perform the desired function. This is different from the model proposed by Zipser and Andersen which describes the behavior of parietal visual neurons in the context of a particular experiment. Since models of data say nothing about underlying mechanisms, they are of limited utility in extrapolating the reaction of real systems to novel experiments [54]. Structural models, like the one described in this paper, in general make much stronger hypotheses about the system in question, and can therefore be more useful for design of future experiments and for understanding of system operation.

Zipser and Andersen's model is based on observations derived from experiments that determine cells' response to single spots of light [30]. While the backpropagation-tuned network may be a good model of the "single spot of light" type of experiment, its accuracy for more complex stimuli (multiple spots or pixels) is not clear. In normal operation, the visual system operates on complex patterns of light. Our model certainly works with arbitrary visual input, and its responses to more complex experimental stimuli can be compared with similar experiments performed on animals to validate or invalidate the model.

Finally, the Zipser and Andersen model was constructed to work over a narrow range of eye positions, and with greatly restricted size (or resolution) of both the retina and the egocentric map. It would be interesting to examine the responses of parietal neurons to changes in eye position that correspond to a few hypercolumns in magnitude, to see how results change as we progress to larger shifts. Even a backpropagation model would no doubt need to include structural elements to accommodate large changes in eye position and large-area, high-resolution retinal and egocentric maps.

4.1 Predictions from the Model

It is instructive to examine how many neurons would be necessary to implement a particular model, to see if that model is plausible. For our model, the number required is dominated by the number of neurons in the final egocentric map. If the topographic relationships within the visual data are to be maintained, then there can be no fewer neurons used in any model. Additionally, even if we assume that the granularity of the egocentric map is such that it represents space at the maximum visual acuity achievable by humans, four stages at most would be required to perform the transform (using the conservative number of 10 cells from one layer providing input to each cell in the next layer). There is no problem fitting such an architecture within the known neuroanatomy of visual areas.

Additionally, since this transform makes such a low demand on the processing power and interconnectivity of neurons, it would not be unreasonable to assume that cells taking part in the transform would also be doing other processing, such as segmenting images based on motion. The result in real terms would be neurons which not only respond to eye position, but also *attributes* of the visual stimulation (for example, the overall optic flow). As previously mentioned, experiments on animals using more complex, realistic stimuli are needed to test the validity of these predictions.

Finally, with respect to the work of Treisman and others in attention, our model predicts that the "master map of locations", within which attention (both internal and motor) operates, is an egocentric map. It favors a model in which this map is placed *after* feature extraction modules.

This model is quite simple, and there is ample room for improvement, both to add more functionality to the model and to fit the model within the context of a more complete visual processing system. Within the scope of this model, it would be interesting to use more realistic neuron models to perform more processing, such as that required to achieve rough segmentation of scenes according to optic flow. Additionally, it would be desirable to integrate binocular disparity computation, so the final output map would be a three-dimensional map of the environment.

This model is meant to be part of a visual processing pathway that performs overall preattentive processing of the entire scene, in order to localize objects and guide direction of attention (whether visual or motor). Thus, the output of our model would be used by higher-level processes which select objects for attentive exploration and construct the connection between object location (computed in parietal cortex) and detailed object shape (computed in temporal cortex) [33]. Performing these operations within a spatial map is essential to our perception of a world with its own objective existence.

References

[1] Jerome A. Feldman. Four frames suffice: a provisional model of vision and space. *Behavior and Brain Sciences*, 8:265–89, June 1985.

[2] E. A. DeYoe and D. C. Van Essen. Concurrent processing streams in monkey visual cortex. *Trends in Neuro Sciences*, 11(5):219–26, 1988.

[3] M. Schlag-Rey, J. Schlag, and B. Shook. Interactions between natural and electrically evoked saccades. i. differences between sites carrying retinal error and motor error signals in monkey superior colliculus. *Experimental Brain Research*, 76:537–47, 1989.

[4] J. Schlag, M. Schlag-Rey, and P. Dassonville. Interactions between natural and electrically evoked saccades. ii. at what time is eye position sampled as

a reference for the localization of a target? *Experiemental Brain Research*, 76:548–58, 1989.

[5] D.L. Robinson and M.E. Goldberg. *The Visual Substrate of Eye Movements*, pages 3–14. *Eye Movements and the Higher Psychological Functions*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

[6] M. Delbrück. *Mind From Matter?* Blackwell Scientific Publications, Palo Alto, 1986.

[7] W.L. Shebilske. *Visuomotor Coordination in Visual Direction and Position Constancies. Stability and Constancy in Visual Perception: Mechanisms and Processes*, John Wiley and Sons, New York, 1977.

[8] I. Bodis-Wollner, M.B. Bender, and S.P. Diamond. *Clinical Observations of Palinopsia: Anomalous Mapping From Retinal to Nonretinal Coordinates of Vision*, pages 659–76. *Sensory Experience, Adaptation, and Perception*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1984.

[9] M.T. Turvey. Contrasting orientations to the theory of visual information processing. *Psychological Review*, 84(1):67–88, 1977.

[10] A. Treisman. Features and objects: the fourteenth bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, 40A(2):201–37, May 1988.

[11] C.W. Eriksen and Y.-Y. Yeh. Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11(5):583–97, 1985.

[12] H. Pashler and P.C. Badgio. Visual attention and stimulus identification. *Journal of Experimental Psychology: Human Perception and Performance*, 11(2):105–21, 1985.

[13] J.R. Bergen and B. Julesz. Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303:696–8, June 1983.

[14] W. Epstein, editor. *Stability and Constancy in Visual Perception: Mechanisms and Processes.* John Wiley and Sons, New York, 1977.

[15] H. Wallach. *On Perception.* Quadrangle, 1976.

[16] T. Kanade. Region segmentation: signal vs. semantics. In *Proceedings of the Fourth International Conference on Pattern Recognition*, pages 95–105, Kyoto, Japan, November 1978.

[17] R. Held and J. Rekosh. Motor-sensory feedback and the geometry of visual space. *Science*, 141:722–3, 1963.

[18] E. von Holst. Relations between the central nervous system and the peripheral organs. *British Journal of Animal Behavior*, 2:89–94, 1954.

[19] M.D. Levine. *Vision in Man and Machine*. McGraw-Hill, 1985.

[20] S. Zucker, A. Rosenfeld, and L.S. Davis. General purpose models: expectations about the unexpected. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence*, pages 716–21, Tbilisi, Georgia, USSR, 1975.

[21] A. Arbib. *The Metaphorical Brain*. Wiley-Interscience, New York, 1972.

[22] Josef Skrzypek. Neural specification of a general purpose vision system. In *Proceedings of ACNN'90, Australian Conference on Neural Networks*, Sydney, Australia, January 1990.

[23] Harry J. Jerison. *Evolution of the Brain and Intelligence*. Academic Press, New York, 1973.

[24] O.-J. Grüsser, A. Krizic, and L.-R. Weiss. Afterimage movement during saccades in the dark. *Vision Research*, 27(2):215–26, 1987.

[25] L. Mays and D. Sparks. *The Localization of Saccadic Targets Using a Combination of Retinal and Eye Position Information*, pages 39–47. *Progress in Oculomotor Research*, Elsevier North Holland, Amsterdam, 1981.

[26] H. von Helmholtz. *HelmholtzUs Treatise on Physiological Optics*. Optical Society of America, 1925.

[27] J.J. Gibson. *The Senses Considered As Perceptual Systems*. Houghton Mifflin, Boston, 1966.

[28] J.H.R. Maunsell and W.T. Newsome. Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10:363–401, 1987.

[29] Richard A. Andersen, Greg K. Essick, and Ralph M. Siegel. Encoding of spatial location by posterior parietal neurons. *Science*, 230:456–8, October 1985.

[30] D. Zipser and R.A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–84, February 1988.

[31] Barry J. Richmond and Michael E. Goldberg. On computer science, visual science, and the physiological utility of models. *The Behavioral and Brain Sciences*, 8:300–1, 1985.

[32] Richard A. Andersen. Head-centered coordinates and the stable feature frame. *The Behavioral and Brain Sciences*, 8:289–90, 1985.

[33] S. Zeki and S. Shipp. The functional logic of cortical connections. *Nature*, 335:311–7, September 1988.

[34] Joaquin M. Fuster. Inferotemporal units in selective visual attention and short-term memory. 1990. forthcoming, Journal of Neurophysiology.

[35] Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240:740–9, May 1988.

[36] B. C. Motter, M. A. Steinmetz, C. J. Duffy, and V. B. Mountcastle. Functional properties of parietal visual neurons: mechanisms of directionality along a single axis. *The Journal of Neuroscience*, 7(1):154–76, January 1987.

[37] Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neuro Sciences*, 414–7, October 1983.

[38] J.D. Schlag. Personal communication. March 1989.

[39] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–4, August 1985.

[40] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[41] A. Treisman and J. Souther. Search asymmetries: a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114:285–310, 1985.

[42] J.D. Schlag and M. Schlag-Rey. *Eye-Movement-Related Neuronal Activity in the Central Thalamus of Monkeys*, pages 169–76. *Progress in Oculomotor Research*, Elsevier North Holland, Amsterdam, 1981.

[43] B. V. Updyke. *Multiple Representations of the Visual Field*, chapter 3, pages 83–101. *Cortical Sensory Organization: Multiple Visual Areas*, Humana Press, Clifton, New Jersey, 1981.

[44] M. H. Pirenne. *Vision and the Eye*. Associated Book Publishers, London, 2nd edition, 1967.

[45] D.H. Hubel. *Evolution of Ideas on the Primary Visual Cortex, 1955-1978: A Biased Historical Account*. *Les Prix Nobel*, Almqvist and Wiksell International, Stockholm, 1981.

[46] Joy Hirsch and Christine A. Curcio. The spatial resolution capacity of human foveal retina. *Vision Research*, 29(9):1095–101, 1989.

[47] Stanley J. Schein and Francisco M. de Monasterio. Mapping of retinal and geniculate neurons onto striate cortex of macaque. *The Journal of Neuroscience*, 7(4):996–1009, April 1987.

[48] José Ambros-Ingerson, Richard Granger, and Gary Lynch. Simulation of paleocortex performs hierarchical clustering. *Science*, 247:1344–8, March 1990.

[49] Eugene Paik and Josef Skrzypek. *UCLA SFINX — Neural Network Simulation Environment.* Technical Report 10, UCLA Machine Perception Laboratory, November 1987.

[50] E. Mesrobian, M. Stiber, and J. Skrzypek. *UCLA SFINX — Structure and Function in Neural Connections.* Technical Report UCLA-MPL-TR 89-8, University of California, Los Angeles Machine Perception Laboratory, November 1989.

[51] J.A. Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–54, 1982.

[52] L. D. Harmon and E. R. Lewis. Neural modeling. *Physiological Reviews*, 46(3):513–91, July 1966.

[53] Ludwig von Bertalanffy. *General System Theory: Foundations, Development, Applications.* Allen Lane, London, 1971.

[54] Joseph J. DiStefano, III and Elliot M. Landaw. Multiexponential, multicompartmental, and noncompartmental modeling. i. methodological limitations and physiological interpretations. *American Journal of Physiology*, 246:R651–4, 1984.