A Model for Parsing, Learning and Recognizing Objects in a Complex Environment

Arnold Trehub
Department of Psychology
University of Massachusetts, Amherst

ABSTRACT

A neuronal model is described that can parse, learn, and recognize objects in a complex visual environment. A computer simulation of the model network was tested with a variety of scenes and exhibits competent performance.

INTRODUCTION

The problem of cognitive adaptation without a prior knowledge base constitutes a ubiquitous and vexing issue in cognitive science. Imagine a person in an absolutely unfamiliar environment, one in which all visual patterns are completely novel. Where would the person look? Since any point of gaze would presumably be no more meaningful than another, how could one parse the scene into objects? How could the objects be learned and committed to memory? This paper presents a computer simulation of a detailed neuronal system that is plausible within biological constraints and can accomplish these fundamental visual-cognitive tasks. The model is composed of several putative neuronal mechanisms proposed in earlier papers (Trehub, 1975, 1977, in press) which have been organized in an integrated system that can deal competently with novel and complex visual environments. The neuronal model will be briefly described and then a computer simulation of the model's behaviour will be presented.

NEURONAL MODEL

Following is an outline of the principal processing elements in the model.

1. Center-surround mechanisms in the retina and lower-level visual nuclei extract simple contours from the light-intensity array.
2. There are cells which integrate contour excitation over small, discrete regions of the entire visual field. These are called flux detectors and serve to drive visual saccades to regions of maximum contour flux.
3. There is a visual field constriction mechanism that can limit the effective stimulus input to an area of variable retinal diameter centered on the foveal axis
4. There is a post-retinal dynamic visual buffer called a retinoid which can translate patterns of retinal stimulation over an egocentric coordinate space. This module locates and positions pattern centroids on a standard reference axis within the visual system.
5. There is an adaptive network called a synaptic matrix which can learn, recognize and image visual patterns.

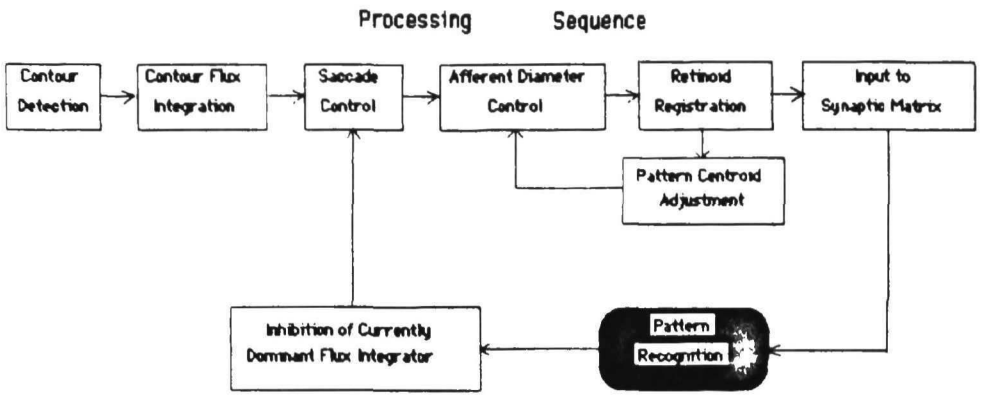

FIGURE 1. Block-flow diagram of processing sequence.

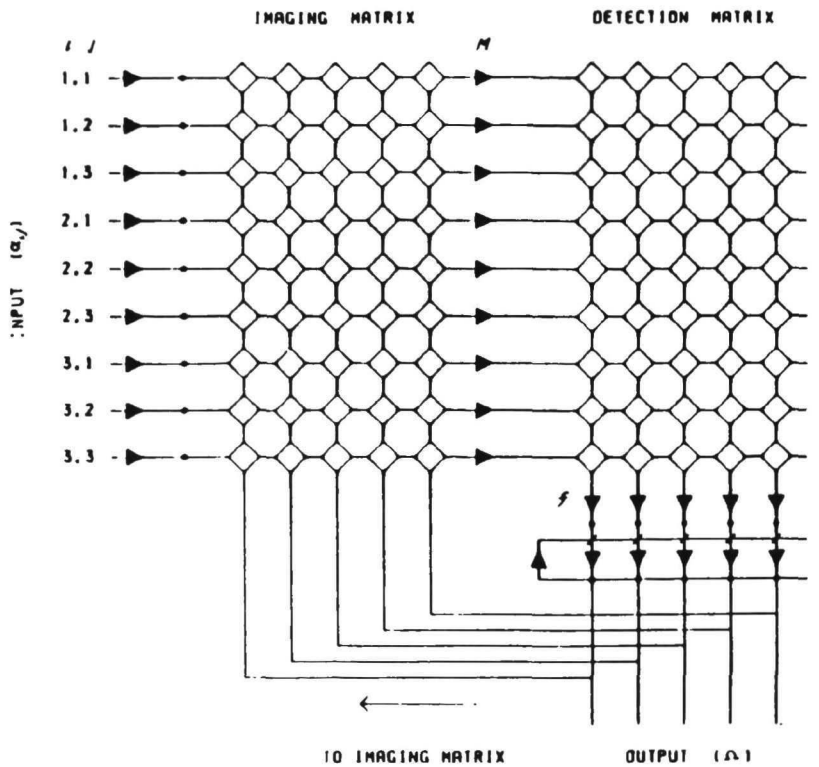

FIGURE 2. Schematic of a synaptic matrix. Afferent inputs from optic tract designated α_{ij}. Mosaic cells designated M. Dots represent fixed excitatory synapses. Short oblique slashes represent fixed inhibitory synapses. Lozenges represent adaptive excitatory synapses. Reset neurons marked (-) generates an inhibitory postsynaptic potential to reset all class cells (Ω) when discharged. Given an arbitrary pattern input, that class cell coupled with the filter cell (f) having the highest product-sum of afferent axon activity ($M_{\alpha ij}$) and corresponding synaptic transfer weights (ϕ_{ij}), will fire first and inhibit the output of all competing class cells.

The diagram shown in Fig. 1 gives a rough representation of the processing sequence. The major modules are outlined below. Space limitations preclude a more detailed presentation of their operating principles which can be found in other publications (Trehub, 1975, 1977, 1985, in press).

Synaptic Matrix. Figure 2 shows a basic version of the neuronal network that has the capability of learning complex retinal input patterns. If a pattern exemplar has been learned, subsequent stimulation by a similar pattern results in the discharge of a particular output cell (class cell) that has been associated with the original exemplar during the learning process. In effect, this cell represent the biological name of its associated pattern. Conversely, the discharge of a class cell alone can generate in an array of mosaic cells the afferent firing pattern (image) initially evoked only by the learned retinal stimulus. Learning occurs in the detection-matrix field when mosaic cells carrying an input pattern fire in virtual coincidence with the discharge of a previously unmodified filter cell, and in the imaging-matrix field when a class cell is fired in coincidence with discharge in the mosaic-cell array. The physical substrate of learning is an adaptive long-term change in the distribution of synaptic transfer weights (ϕ) on the dendrites of filter cells and mosaic cells.

Retinoids. The neuronal structure shown in Fig. 3 is a post-retinal mechanism called a retinoid because it represents visual space and projects afferents to the mosaic-cell array. This module may be thought of as a visual scratch-pad with phasic and dynamic content. The medium of storage is assumed to be a retinotopically organized sheet of excitatory autaptic neurons. Cells of this type have at least one of their axon collaterals in recurrent excitatory synapse with their own cell body or dendrite (Shepherd, 1974).

If there is a pattern of excitation evoked on a retinoid, this captured pattern can be spatially translated in any direction by appropriate pulses from the shift command cells. For example, each pulse from the shift-right line will transfer standing activity from each active autaptic cell to the adjacent autaptic cell on its right and, at the same time, erase activity in the previously active cell (the donor cell) unless that cell is also receiving transfered excitation from an autaptic cell to its immediate left. The more rapid the pulses, the more rapid will the pattern move; the longer the pulse train is sustained, the greater will be the distance over which the pattern is moved. Appropriate sequences of shift right/left, shift up/down, can move the pattern of cell activity to any position on the retinoid surface.

Imagine the retinoid as a quadrantally organized surface, with each quadrant receiving retinotopic afferents from its respective retinal quadrant. If the excitation of a standing pattern is summed independently over each quadrant, and if the relative magnitudes of the summed discharges are used to drive either the position of the eye or the shift control cells in the retinoid, then we have a neuronal mechanism which can align the centroid of any retinal stimulus with the central axis of retinoid space (Trehub, 1985). We define the normal foveal axis as that axis corresponding to the line of sight of the fovea when the eyes are straight ahead, the head unturned, and the shoulders square with the body. It is assumed that the central axis in retinoid space corresponds with the normal foveal axis.

The quandrantal summation fields for retinoid output are abbreviated as follows:

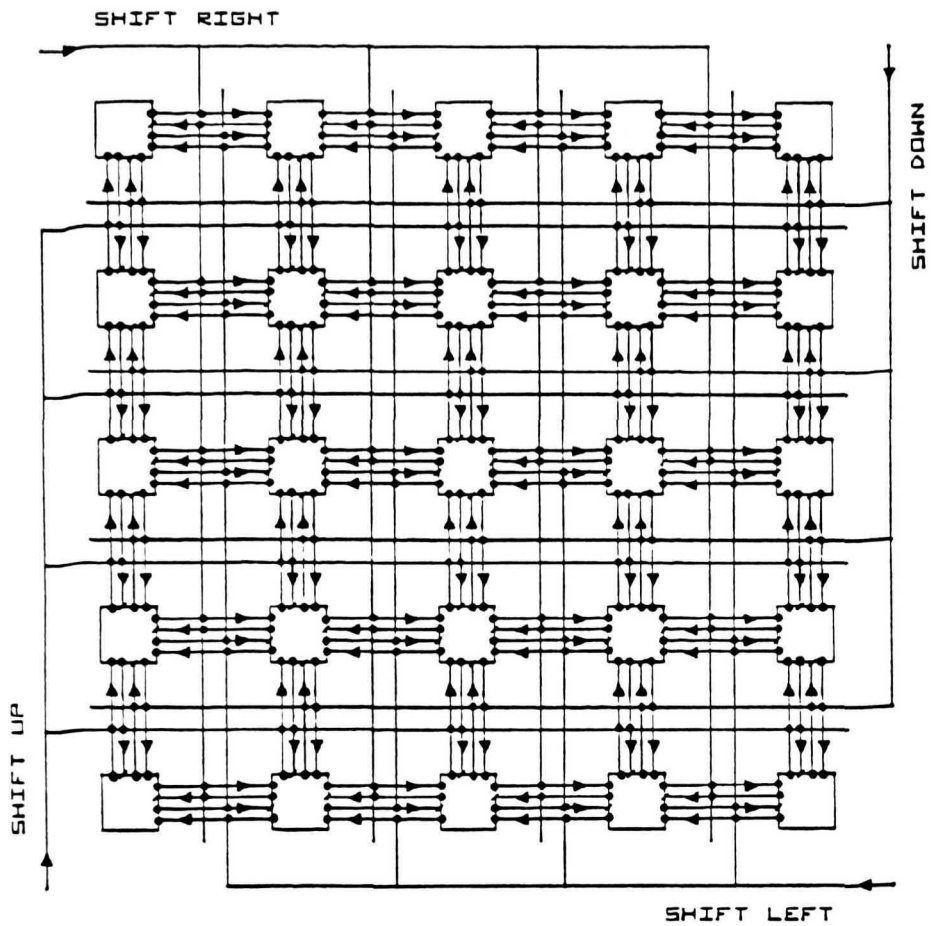

FIGURE 3. Translation retinoid. Large squares represent autaptic cells serving short-term memory. Small filled triangles represent interneurons. Shift-control cells designated by direction of effect.

LF = output from left retinoid field.
RF = output from right retinoid field.
TF = output from top retinoid field.
BF = output from bottom retinoid field.

If the difference between total output in RF-LF and TF-BF respectively were to drive an eyeball in the direction of the greater excitation in the hemi fields defined by the two orthogonal axes, the fovea would hunt until it targeted the contour centroid of any stimulus pattern presented to the retina. Alternatively, if the point of eye fixation does not change, then a pattern with a parafoveal centroid can be translated over the retinoid surface so that its centroid falls on the normal foveal axis of the retinoid. This is done by using the hemifield mismatches to drive the

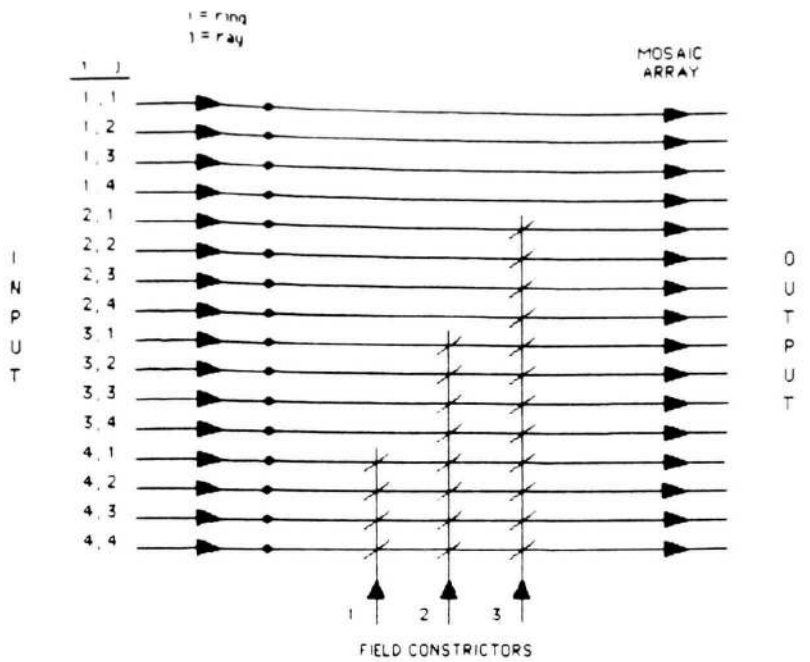

FIGURE 4. Controls for constricting effective visual field. Discharge of constrictor neuron 1 blocks input from ring 4 (outer ring); discharge of constrictor 3 blocks input from rings 4, 3, 2, restricting input to ring 1, the innermost ring of afferents.

shift-control cells so that excitation is balanced over retinoid quadrants.

Field Constrictor. It is possible to devise a number of different coordinate representations for retinotopic indexing, but I have found a ring-ray representation to be particularly useful and efficient. In this scheme, receptor cells in the retina and their associated afferent projections are indexed with respect to the central foveal axis in terms of their locations on imaginary concentric rings (i) centering on the axis, and imaginary rays (j) projecting from the axis and intersecting all rings. This retinal organization easily lends itself to CNS control of the afferent field aperture. Figure 4 shows how inhibitory neurons can impinge successively on entire rings of mosaic cells to constrict the diameter of the effective visual field.

COMPUTER SIMULATION

A 22x22 cell retina and the neuronal mechanisms outlined above were simulated in a digital computer. Indoor (near) and outdoor (far) environments were created in sketch-to-pixel conversions, and these environments were presented to the simulated visual system for parsing, learning, and object recognition.

At the start of each scene-parsing operation, the model first fixated on the retinotopic locus of the flux detector with maximum output, then

the afferent field aperture closed to the fully constricted state which was arbitrarily set at six retinal units in width and height. The fully expanded afferent aperture was limited to 22x22 retinal units. Whenever the visual aperture reached the state of full expansion, the excitation pattern on the retinoid was gated to the synaptic matrix for recognition (and learning if the pattern was incorrectly identified). Starting error tolerance was set at three units for quadrantal disparity over either the horizontal or vertical axes. At any fixed aperture, if error tolerance was exceeded on a given axis, the retinoid pattern was shifted in the appropriate direction to reduce hemifield disparity on that axis. When pattern position satisfied error tolerance for one axis, the pattern was shifted on the other axis, unless it was already within tolerance. If, now, shifting the image on the second axis resulted in an unacceptable error on the first, error tolerance was relaxed one unit. Whenever the pattern was brought within axial tolerance for both horizontal and vertical disparities, the afferent aperture expanded one unit and the process was repeated until full aperture was achieved. This operation was assumed to involve an expenditure of processing effort, and if a retinoid shift of nine units on any axis did not bring its disparity within tolerance limits, the system stopped trying at its current fixation and initiated a saccade to the next highest flux region.

In its initial state, the neuronal system is presented with a random visual pattern and taught to call this pattern "RANDOM". This simply means that one filter cell in the detection matrix and a spatially correlated array of mosaic cells in the imaging matrix have been synaptically tuned to the random exemplar.

The first "natural" environment learned was an outdoor scene consisting of trees, a house, several animals, a building, a car, and the outline of distant hills. Since the simulation does not incorporate mechanisms of visual accommodation or stereopsis (see Trehub, 1978), the operator is asked by the model to provide a rough estimate (in feet) of its viewing distance from the major elements of the scene. The operator estimates the distance as 200 feet and provides this information to the network. Parsing then proceeds according to the principles discussed above. After a pattern has been fixated and registered on the retinoid, it is passed to the synaptic matrix where it is identified and named as "RANDOM" because, in its utterly naive condition, this is its only available response. The model then asks the operator to inform it if the response is right or wrong. Let us say that the object it has happened to parse is a house or part of a house; then it is told that the response is wrong. At this point, the model changes the synaptic weights on a previously unmodified filter cell in accordance with the excitation pattern on its mosaic-cell array and the learning equation. It should be noted here that if there were no operator to inform the system about the correctness of its response, low frequency discharge of its filter cells can provide a signal that the current stimulus is novel, triggering the automatic learning of the novel object (Trehub, 1977). After the filter cell has been tuned to the stimulus, the model asks for a name to be associated with the class cell which is coupled to the just-modified filter cell. The operator then provides the appropriate name "HOUSE". This name then becomes part of a neuronal lexicon in which it is connected with the filter-cell-class-cell couplet which has just learned the exemplar of a house. The model then parses another object and if its recognition response is correct, parsing continues; if incorrect,

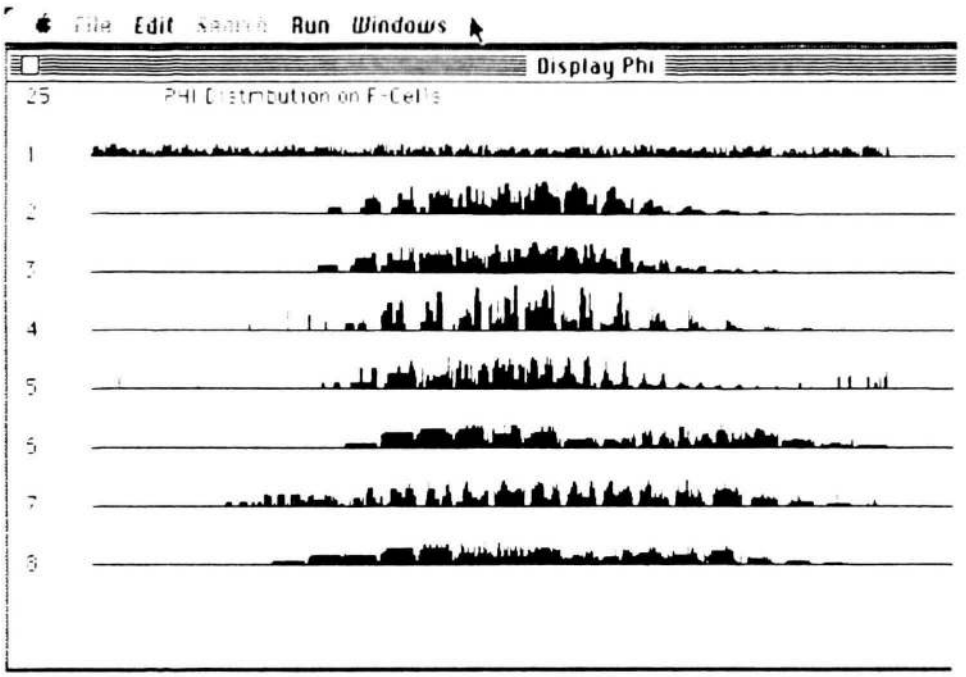

FIGURE 5. Examples of the distribution of synaptic transfer weights on dendrites of filter cells which have learned visual patterns. Each point on the dendritic line represents a particular synaptic location. Amplitude of each vertical line represents relative magnitude of transfer weight for that synapse. The objects learned by the cells shown are as follows: (1) a random visual pattern; (2) a car; (3) a different car; (4) an animal; (5) a different animal; (6) a building; (7) a house; (8) a different building.

the new object is learned (synaptic modification of another available filter cell, etc) and scene processing continues until a preset number of saccades are made, during which objects are fixated, translated to the normal foveal axis, recognized, and learned if necessary.

The second environment learned was a desktop with a book, telephone, ashtray, pencil, and bookmark. The viewing distance, in this case, was estimated to be five feet. Parsing and learning the objects in this scene then proceeded as in the outdoor environment. Variations of both kinds of environments were created and exposed to the model until a total of 25 exemplars of objects in these scenes were learned together with their appropriate names. Examples of synaptic transfer-weight (ϕ) distributions on filter-cell dendrites for the first eight patterns learned are shown in Fig. 5. The selectivity of recognition response is determined by the differences among such ϕ-distributions over the population of filter cells in the detection matrix. As the repertoire of exemplar-tuned filter cells increased, the frequency of recognition errors decreased.

Shown in Fig. 6 is a run of the simulation printed directly from the computer's CRT. In this case, the model was "looking at" an unfamiliar outdoor scene, in that all the patterns in the environment were new exemp

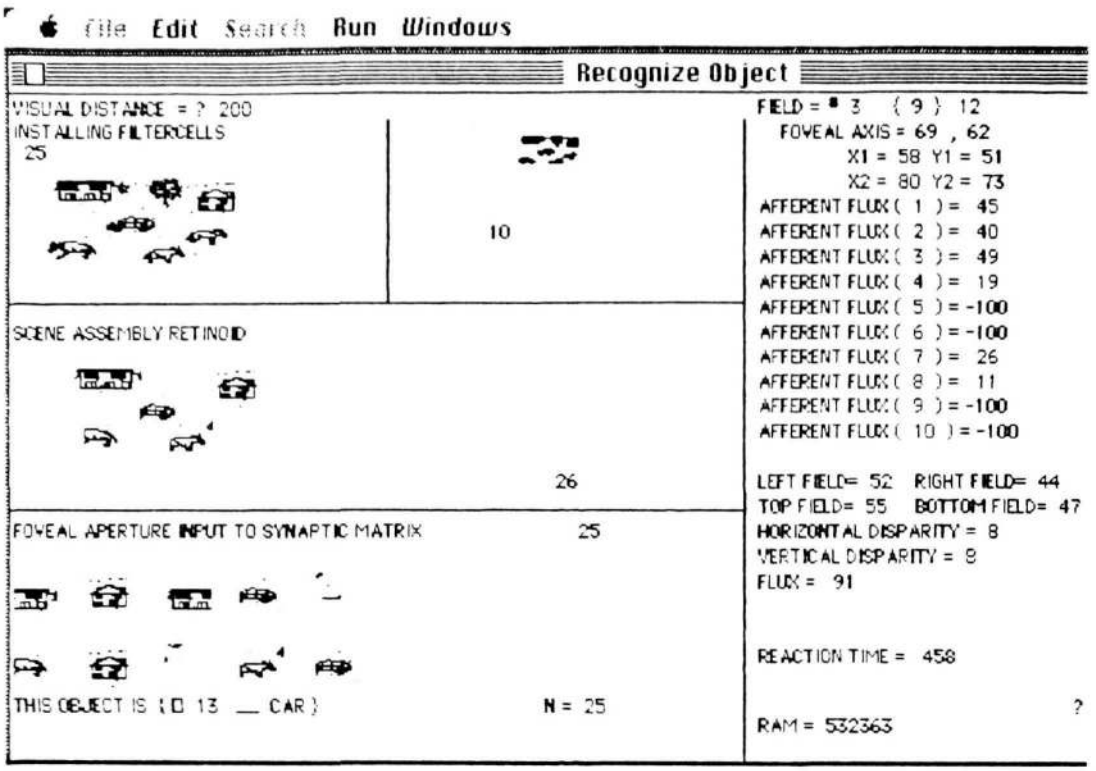

FIGURE 6. Passive recognition. Model's responses to outdoor scene. Top left frame is the scene presented. Bottom left shows objects parsed and recognized. All objects were correctly identified. Middle left is the visual reconstruction of the scene on a retinoid surface on the basis of the disparate fixations and parsings.

lars of previously learned objects and their locations were different. Figure 7 is a similar printout of a situation in which the model is "asked" to find named objects on a cluttered desktop. Here parsing and recognition is made even more difficult by the fact that a bookmark has been placed on the book and a substantial part of the book is covered with a sheet of paper. It has been conjectured that occlusions of this kind as well as the conjunction of nearby objects would make it impossible for template/filter models to operate properly (Pinker, 1984). The successful performance of the model described here suggests that the conjecture is incorrect.

In summary, computer simulation of an explicit and biologically plausible neuronal model demonstrates that a visual system that integrates (a) contour flux detection, (b) flux-driven saccades, (c) control of afferent-field aperture, (d) a retinoid for pattern centroid alignment, and (e) a synaptic matrix for pattern learning can start without an initial store of world knowledge, be exposed to novel and complex scenes, and build an appropriate knowledge base. Confronted with a rich, new visual environment, it isolates objects, learns them, and recognizes similar objects in other environments.

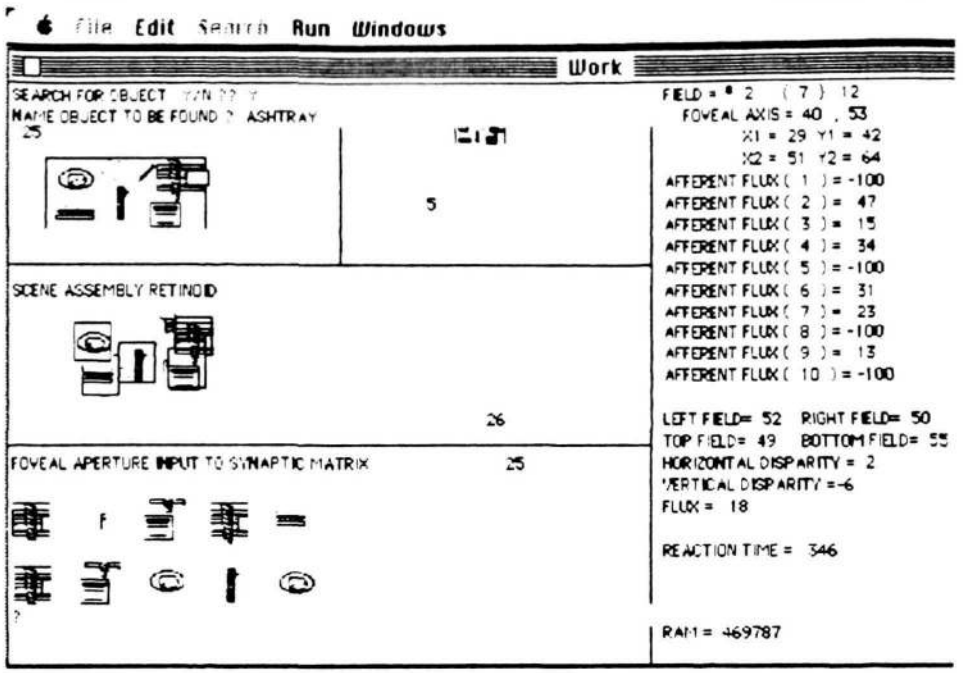

FIGURE 7. Active search and recognition. Model's responses to desktop scene. Small rectangular frame around parsed objects on the scene assembly retinoid indicates that a searched-for object has been found.

REFERENCES

Pinker, S. (1984). Visual cognition: An introduction. Cognition, 18, 1-63.

Shepherd, G. M. (1974). The Synaptic Organization of the Brain. New York: Oxford University Press.

Trehub, A. (1975). Adaptive pattern processing in the visual system. International Journal of Man-Machine Studies, 7, 439-446.

Trehub, A. (1977). Neuronal models for cognitive processes: Networks for learning, perception and imagination. Journal of Theoretical Biology, 65, 141-169.

Trehub, A. (1978). Neuronal model for stereoscopic vision. Journal of Theoretical Biology, 71, 479-486.

Trehub, A. (1985). A confusion matrix for hand-printed alphabetic characters: Testing a neuronal model. Eighth Symposium on Quantitative Analysis of Behavior, at Harvard. (Cambridge, Massachusetts).

Trehub, A. (In press). Visual-cognitive neuronal networks. In M. A. Arbib and A. R. Hanson (Eds.), Vision, Brain, and Cooperative Computation. Cambridge: MIT Press.