A Model of Visual Perception and Recognition Based on Separated Representation of "What" and "Where" Object Features

Ilya A. Rybak
University of Pennsylvania, Philadelphia, PA 19104
ilya@shiva.seas.upenn.edu

Abstract

In the processes of visual perception and recognition human eyes actively select essential information by way of successive fixations at the most informative points of the image. So, perception and recognition are not only results or neural computations, but are also behavioral processes. A behavioral program defining a scanpath of the image is formed at the stage of learning (object memorizing) and consists of sequential motor actions, which are shifts of attention from one to another point of fixation, and sensory signals expected to arrive in response to each shift of attention.

In the modern view of the problem, invariant object recognition is provided by the following: (i) separated processing of "what" (object features) and "where" (spatial features) information at high levels of the visual system; (ii) mechanisms of visual attention using "where" information; (iii)representation of "what" information in an object-based frame of reference (OFR).

However, most recent models of vision based on OFR have demonstrated the ability of invariant recognition of only simple objects like letters or binary objects without background, i.e. objects to which a frame of reference is easily attached. In contrast, we use not OFR, but a *feature-based frame of reference* (FFR), connected with the basic feature (edge) at the fixation point. This has provided for our model, the ability for invariant representation of complex objects in gray-level images, but demands realization of behavioral aspects of vision described above.

The developed model contains a neural network subsystem of low-level vision which extracts a set of primary features (edges) in each fixation, and high-level subsystem consisting of "what" (Sensory Memory) and "where" (Motor Memory) modules. The resolution of primary features extraction decreases with distances from the point of fixation. FFR provides both the invariant representation of object features in Sensory Memory and shifts of attention in Motor Memory. Object recognition consists in successive recall (from Motor Memory) and execution of shifts of attention and successive verification of the expected sets of features (stored in Sensory Memory). The model shows the ability of recognition of complex objects (such as faces) in gray-level images invariant with respect to shift, rotation, and scale.

Introduction

It is known that in the processes of visual perception and recognition, human eyes move and successively fixate at the most informative points of the image (Noton & Stark, 1971; Yarbus, 1967). During these processes, the eyes actively perform problem-oriented selection of information from the visible world under the control of mechanisms of visual attention (Burt, 1988; Julesz, 1975; Neisser, 1967; Shiffrin & Schneider, 1977; Triesman & Gedal, 1980; Yarbus, 1967). Thus, *visual perception and recognition are not only a result of computations that are performed by neural networks of the visual system but are also behavioral processes.*

An important component of behavior is the behavioral "program" which is formed for each new situation during learning. A behavioral program consists of sequential motor actions and the sensory signals which are expected to arrive in response to execution of each motor action. Behavior may be subdivided into two basic stages: the

stage of selection of the appropriate behavioral program (making a hypothesis about the situation), and the stage of its execution. The successive matching of expected sensory signals with current ones after each motor action plays a major role in the execution stage.

The idea that perception and recognition of visual images are behavioral processes and possess these properties was proposed by Noton and Stark (1971). They carried out research devoted to comparing individual scanpaths of human eye movements in two phases: when an object was being memorized and when it was being recognized. They have found that these scanpaths are topologically similar and have suggested that each object is memorized and stored in memory as a sequence of object features and eye movements required to reach the next feature along the corresponding scanpath. Based on this idea, it is possible to consider a scanpath of an image, consisting of sequential eye movements, to be a behavioral program. *The whole image in this case must be represented and stored in memory as a sequence of image fragments processed at the fixation points along a scanpath. The process of recognition consists of successive eye movements recalled from motor memory and successive verification of the expected image fragments.*

There is neuro-anatomical and psychological evidence that high levels of the visual system contain two major pathways for visual processing. One pathway, called the "where" pathway, leads dorsally to the parietal cortex, and is involved in processing and representation of spatial information (spatial locations and relationships). The other pathway, called the "what" pathway, leads ventrally to the inferior temporal cortex and deals with processing and representation of object features (Kosslyn et al., 1990; Underleider & Mishkin, 1982; Van Essen, 1985). Many researchers believe that invariant object recognition in human vision is provided by the following: 1) separated processing of "what" (object features) and "where" (spatial features) information at high levels of the visual system; 2) mechanisms of visual attention using "where" information; 3)representation of "what" information in an *object-based frame of reference* (Hinton & Lang, 1985; Marr, 1982; Palmer, 1983).

However, most recent models of vision based on the object-based frame of reference have demonstrated the ability of invariant recognition of only simple objects like letters or binary objects without background, i.e. objects to which a frame of reference is easily attached (Ahmad, 1992; Carpenter, Grossberg, & Lesher, 1992; Olshausen, Anderson, & Van Essen, 1992; Otto et al., 1992; Rueckl, Cave, & Kosslyn, 1989). In contrast, we believe that the human visual system uses not an object-based, but rather a *feature-based frame of reference*, connected with the basic feature (local edge) at the fixation point. However, the realization of this idea demands taking into account the data on high-level visual processing and behavioral aspects described above.

The Model

A functional diagram of the model is shown in Fig. 1. The attention window provides a primary transform of the processed image into a "retinal image" at some point of fixation. This primary transform is a nonlinear transformation of the image that provides decreasing resolution from the center of the attention window to its periphery. It simulates the decrease of resolution from the fovea to the retinal periphery in the cortical map of the retinal image.

The "retinal image" extracted from the attention window goes to a module for primary feature extraction that performs the function of the primary visual cortex. This module consists of neurons with orientationally selective receptive fields tuned to different orientations of local edge segments. Neurons with overlapping receptive fields but with different orientation tuning interact competitively owing to strong reciprocal inhibitory interconnections. The orientation tuning of the winning neuron at each location defines the orientation of the edge at that point.

In each fixation (position of the attention window), oriented edge segments are extracted at the fixation point (center of the attention window) and at a number of

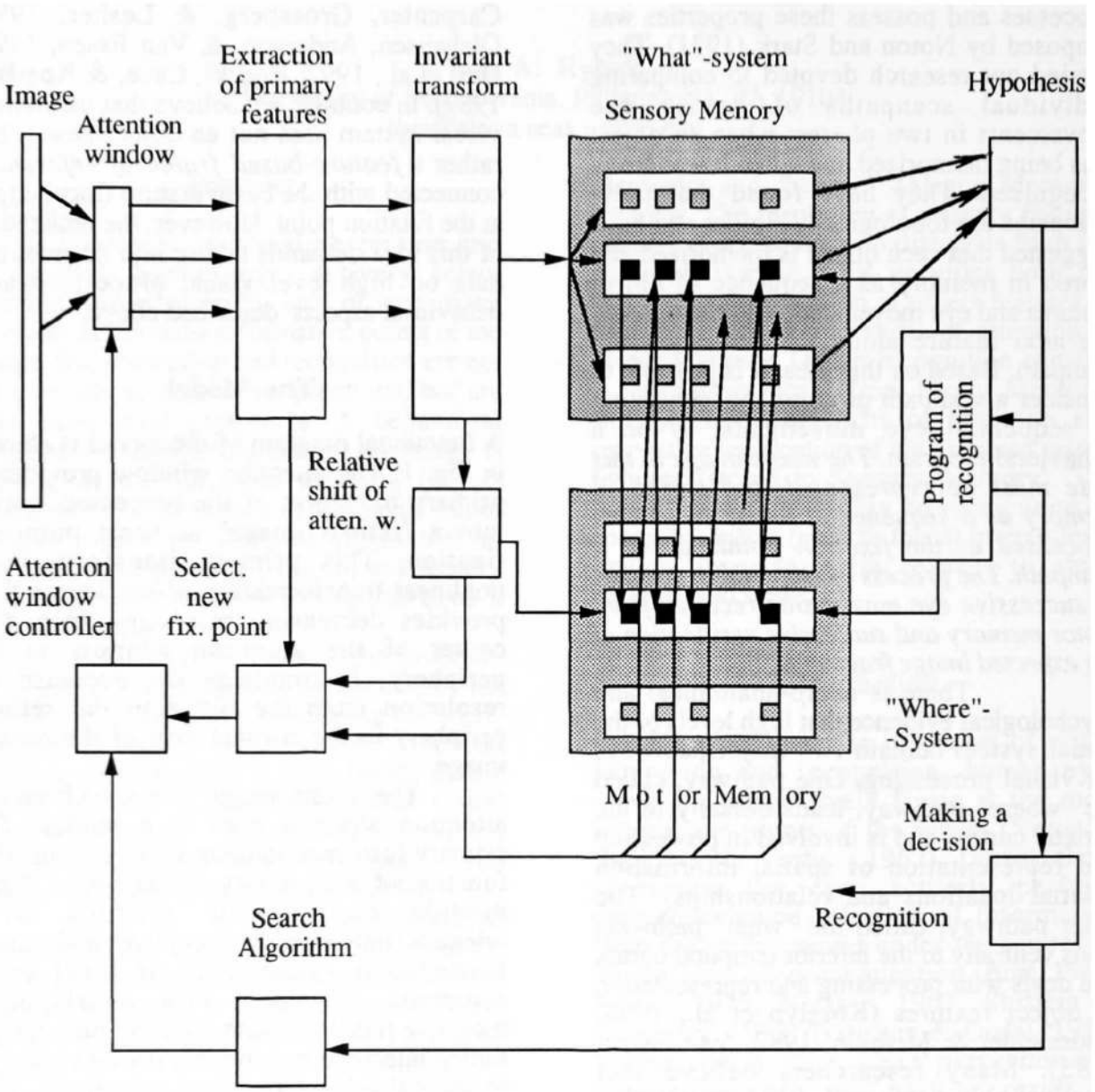

FIGURE 1. Functional diagram of the model.

other points located at specific positions in the retinal image. Let us call the central point the "basic" point and the other ones "context" points. Also let us call the edge segments at the basic and context points "basic edge segment" and "context edge segments", respectively. The positions of the context points are fixed with respect to the fovea and have a constant angle step, but their distances from the center, depend on the resolution in the central area of the attention window. Thus, in the output of the subsystem for primary feature extraction, there is a set of oriented edge segments (the basic and several context ones) that characterize the image fragment at the given point of fixation with decreased resolution toward the periphery of the attention window.

The modules described above form a low-level subsystem of the model. The next module may be considered to be a middle level of processing. It transforms the set of primary features, oriented edge segments, into invariant second-order features. The method of this transformation is based on the idea that the coordinate system (frame of reference) is attached in each fixation to the basic edge segment in the center of the retinal image (attention window).

The *relative orientations of context edges* and their *relative angle positions* with respect to the coordinate system connected with the basic edge segment are considered to be second-order features.

The functioning of the high-level subsystem and the whole model may be considered in three different modes: the mode of image memorizing, the mode of image search, and the mode of image recognition.

In the mode of image memorizing, the image is processed at successively selected fixation points of attention. At each fixation point, the set of oriented edge segments (the basic and several context ones) is extracted from the retinal image (attention window). Then, this set is transformed into vectors of invariant second-order features. These vectors are memorized in the neural network which plays the role of long-term Sensory Memory ("what"- system). The position of the next fixation of attention is selected from the set of context points and is represented with respect to a coordinate system defined by the basic edge segment. Each relative shift of attention is memorized (as an invariant spatial feature) in Motor Memory ("where"-system). A special module then computes the new fixation point position in absolute coordinates, thereby controlling the shift of attention to a new fixation point via the attention window controller, which plays the role of the oculomotor system. As a result of the memorizing mode, a sequence of image fragments will be memorized in the "what"-system (Sensory Memory), but a sequence of shifts of attention will be memorized in the "where"-system (Motor Memory). These two sequences of two types of memory alternate in a chain of elements that form the behavioral program of viewing (a program of recognition) for the memorized image.

In the mode of image search, the image is scanned by the attention window under the supervision of a search algorithm. In each position of the attention window, the current fragment from the attention window is compared with all fragments of all objects memorized in Sensory Memory. Scanning of the image in the mode of image search continues until a fragment similar to some memorized fragment is found at some fixation point. When such a fragment is found, a hypothesis of the image is generated and the model turns to the recognition mode.

In the mode of recognition, the program is executed by the way of consecutive shifts of the attention window (controlled from Motor Memory) and consecutive verification of similarity of the current image fragments with fragments stored in Sensory Memory. A scanpath of viewing in the recognition mode sequentially reproduces the scanpath of viewing in the memorizing mode. If a series of successful matches occurs, a decision is made that the object is recognized. If it does not, the model returns to the mode of image search.

The test experiments have shown that the developed model is able to recognize complex gray-level images invariant with respect to shift, rotation, and scale. An example is shown in Fig. 2. In Figure 2a, the scanpath shown was formed in the memorizing mode. Figure 2 b and c shows the scanpaths that took place during the process of successful recognition of the test images which were rotated (b) or reduced in scale (c).

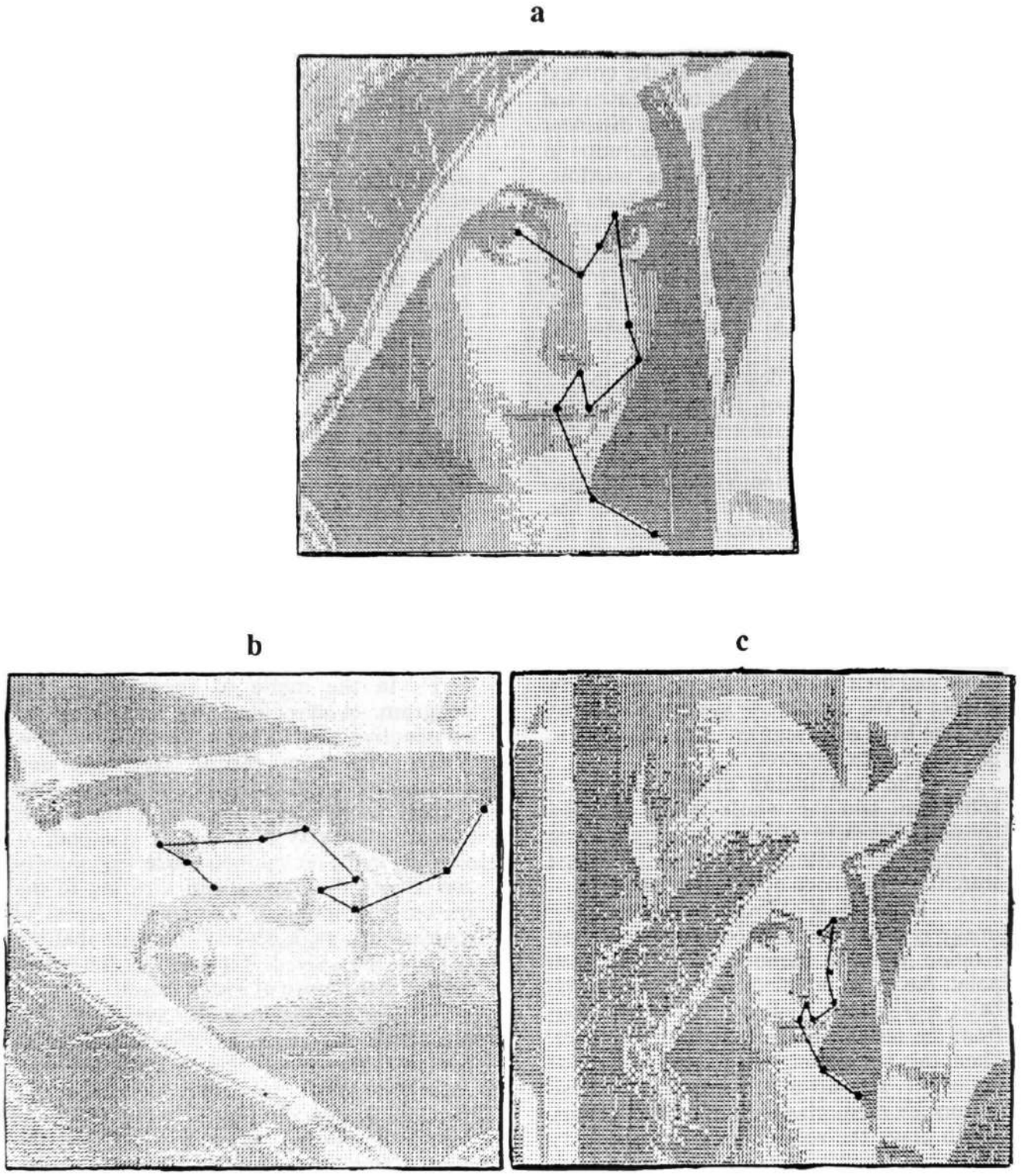

FIGURE 2. Excmple of recognition of test image.

References

Ahmad, S. 1992. VISIT: A neural model of covert visual attention. In *Advances in Neural Information Processing Systems* 4, Morgan Kaufmann , 1992.

Burt, P.J. 1988. Smart sensing within a pyramid vision machine. *Proceedings of the IEEE* 76: 1006-1015.

Carpenter, G.A., Grossberg, S., and Lesher, G.W. 1992.A what-and-were neural network for invariant image preprocessing. In *Proceedings of International Joint Conference on Neural Networks* 3, 303-308. Baltimore.

Hinton, G.E. and Lang, K.J. 1985. Shape recognition and illusory conjunctions. In *Proceedings of Ninth International Joint Conference on Artificial Intelligence*. Los Angeles.

Julesz, B. 1975. Experiments in the visual perception of texture. *Scientific American* 232: 34-43.

Kosslyn, S.M., Flynn, R.A., Amsterdam, J.B., and Wang, G. 1990. Components of high-level vision: a cognitive neuroscience analysis and account of neurological syndromes. *Cognition* 34: 203-277.

Marr, D. 1982. *Vision*. N.Y. W.H.Friman.

Neisser, V. 1967.*Cognitive Psychology*. N.Y. Appleton.

Noton, D. and Stark, L. 1971. Scanpaths in eye movements during pattern recognition. *Science*. 171: 72-75.

Olshausen, B., Anderson, C., and Van Essen, D. 1992. A neural model of visual attention and invariant pattern recognition, CNS Memo 18. Caltech. Pasadena.

Otto, I., Grandguillaume, P., Boutkhil, L., and Burnod, Y. 1992. Direct and indirect cooperation between temporal and parietal networks for invariant visual recognition. *Journal of Cognitive Neuroscience* 4: 35-57.

Palmer, S.E. 1983. The psychology of perceptual organization: a transformational approach. *Human and Machine Vision*, Academic Press.

Rueckl, J.G., Cave, K.R., and Kosslyn, S.M. 1989. Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience* 1: 171-186.

Shiffrin, R.M., and Schneider, W. 1977. Controlled and automatic human information processing. 2. Perceptual learning, automatic attending and a general theory. *Psychological Review* 84: 1270-1290.

Triesman, A.M., and Gedal, G. 1980. A feature integration theory of attention. *Cognitive Psychology* 12: 97-136.

Ungerleider, L.G., and Mishkin, M. 1982. Two cortical visual systems. *Analysis of Visual Behavior* . Cambridge. MIT Press.

Van Essen, D. 1985. Functional organization of primate visual cortex. *Cerebral Cortex* 3. N.Y. Plenum.

Yarbus, A.L. 1967. *Eye Movements and Vision* . N.Y. Plenum.