Neurobiology of Attention

Edited by Laurent Itti, Geraint Rees, and John K. Tsotsos

CHAPTER

108

Attention-Guided Recognition Based on "What" and "Where" Representations: *A Behavioral Model*

Ilya A. Rybak, Valentina I. Gusakova, Alexander V. Golovan, Lubov N. Podladchikova and Natalia A. Shevtsova

ABSTRACT

We describe the model of attention-guided visual perception and recognition previously published in *Vision Research* (Rybak et al., 1998). The model contains (1) a low-level subsystem that performs a fovealike transformation and detection of primary features (edges) and (2) a high-level subsystem that includes separated "what" (sensory memory) and "where" (motor memory) subsystems. In the model, image recognition occurs under top-down control of visual attention during the execution of a behavioral recognition program formed during the primary viewing of the image. The recognition program contains both programmed movements of an attention window (stored in the motor memory) and predicted image fragments (stored in the sensory memory) for each consecutive fixation. The model shows the ability to recognize complex images (e.g., faces) invariantly with respect to shift, rotation, and scale.

I. INTRODUCTION

During visual perception and recognition, human eyes move and successively fixate at the most informative parts of the viewed image or scene. Because the density of photoreceptors in the retina decreases from in the central area (fovea) to the periphery, the resolution of image representation in the visual cortex is the highest for the part of the image projected onto the fovea and decreases rapidly with the distance from the fovea projection. The major function of visual attention is to actuate and control eye movement and hence to perform a problem-oriented selection and processing of information from the visible world (Burt, 1988; Neisser, 1967; Noton and Stark, 1971; Posner and Presti, 1987; Treisman and Gelade, 1980; Yarbus, 1967). Despite a relatively low resolution of image representation in the retinal periphery, it provides important information that is used by the attention mechanisms for selecting the next eye position and directing the consecutive eye movements. This leads to the view that visual perception and recognition are actually behavioral processes that probably cannot be completely understood in the limited frames of neural computations without taking into account the behavioral and cognitive aspects of perception and the role of attention in visual perception and recognition.

A. A Behavioral Paradigm in Visual Perception and the Role of Visual Attention

From the behavioral point of view, an internal representation (model) of new circumstances is formed in the brain during conscious observation and active examination. The active examination is aimed toward the finding and memorizing of functional relationships between the applied actions and the resulting changes in sensory information. An external object becomes

known and may be recognized when the system is able to subconsciously manipulate the object and predict the object's reactions to the applied actions. According to this paradigm, the internal object representation contains chains of alternating traces in motor and sensory memories. Each of these chains reflects an alternating sequence of elementary motor actions and sensory (proprioceptive and exteroceptive) signals that are expected to arrive in response to each action. The brain uses these chains as behavioral programs in subconscious behavioral recognition when the object is (or is assumed to be) known. This behavioral recognition has two stages: conscious selection of the appropriate behavioral program and the subconscious execution of the program. Matching the expected (predicted) sensory signals to the actual sensory signals, arriving after each motor action, is the essential procedure in the program execution.

This behavioral paradigm has been formulated and developed in the context of visual perception and recognition in the series of conceptually significant works (Didday and Arbib, 1975; Noton and Stark, 1971; Kosslyn et al., 1990; Yarbus, 1967). Using Yarbus's approach, Noton and Stark (1971) compared the individual scanpaths of human eye movements in two phases: during memorizing and during subsequent recognition of the same image. They found these scanpaths to be topologically similar and suggested that each object is memorized and stored in memory as an alternating sequence of object features and eye movements required to reach the next feature. The results of Noton and Stark (1971) and Didday and Arbib (1975) prompted the consideration of eye movement scanpaths as behavioral programs for recognition. The process of recognition was supposed to consist of an alternating sequence of eye movements (recalled from the motor memory and directed by attention) and verifications of the expected image fragments (recalled from the sensory memory).

Ungerleider and Mishkin (1982), Mishkin et al. (1983), and Kosslyn et al. (1990) presented neuroanatomical and psychological data complementary to this behavioral concept. It was found that the higher levels of the visual system contain two major pathways for visual processing, called the "where" and "what" pathways. The "where" pathway leads dorsally to the parietal cortex and is involved in processing and representing spatial information (spatial locations and relationships). The "what" pathway leads ventrally to the inferior temporal cortex and deals with processing and representing object features (Kosslyn et al., 1990; Mishkin et al., 1983; Ungerleider and Mishkin, 1982). The behavioral concept joined with this neuroanatomical theory provides (1) the explicit functional coupling between the low-level vision (foveal structure of the retinocortical projection, orientation selectivity in the visual cortex, etc.) and the high-level brain structures involved in visual perception and recognition and (2) the clear functional role of visual attention in the coupling the low- and high-levels of the visual system.

B. Image Features, Invariant Representation, and Frames of Reference

Beginning with the classic Hubel and Wiesel (1962) work, neurophysiological studies have demonstrated that neurons in the primary visual cortex can detect elementary image features such as local orientations of line segments or edges. Therefore, most theories of vision assume that the visual system detects simple features (e.g., local line segment or edge orientation but not spatial relations between them) at the preattentive stage and uses some attention mechanisms of a serial type to bind the simple features into more complex shape features (Neisser, 1967; Treisman and Gelade, 1980).

One key issue of visual recognition is the mechanism used for invariant image representation. Marr (1982), Palmer (1983), Hinton and Lang (1985), and others assumed that the visual system uses an object-based frame of reference attached to the center of the object. However, the object-based reference paradigm has several significant disadvantages. First, this paradigm presumes that the object is isolated and does not have a complex background. Second, if a part of an object is missing or occluded, or an additional part is present, the center of the object may be shifted, that makes it difficult to recognize the object. In addition, this paradigm presumes that the object is small and simple enough to be recognized during one gaze fixation. As a result, previous models of recognition that used the object-based frame of reference could demonstrate invariant recognition of only simple objects (letters, binary objects without background, etc.) to which such a frame of reference is easily attached.

In the model reviewed here (Rybak et al., 1998), a spatial pattern of edges is extracted from the retinal image at each fixation. These edges are considered as elementary or first-order features of the image. Image representation at the fixation point is based on the assumption that the first-order features (edges) extracted from the retinal periphery perform two distinct functions. One function of the peripheral edges is to provide potential targets for the next gaze fixation (which was used in most active vision models; see, for

example, Burt, 1988). The other function of peripheral edges, as suggested in the reviewed model, is to provide a context (context features) for the basic feature (edge) in the center of the fovea.

In the reviewed model, the relative orientations of context edges and their relative angular locations with respect to the basic edge at each point of fixation are used as second-order invariant features of the image. Thus, instead of the object-based frame of reference, the model uses a feature-based frame of reference attached to the basic edge at the fixation point. Because both the retinal images at the fixation points and the sequential shifts of the fixation point are represented in this invariant form, the entire image is invariantly represented in the memory of the system. Moreover, the feature-based frame of reference coupled with the multiplicity of fixation points along the scanpath may allow the system to recognize an image from its part (from a fraction of the scanpath belonging to this part) when the image is partly perturbed or the object of recognition in the image is occluded.

II. A MODEL OF ATTENTION-GUIDED RECOGNITION BASED ON "WHAT" AND "WHERE" REPRESENTATIONS

A. Model Description

A functional diagram of the model is shown in Fig. 108.1. The attention window (AW) performs a primary transformation of the image into a retinal image at the fixation point (see Figs. 108.2 and 108.3B). This transformation provides a decrease in resolution of the retinal image from the center to the periphery of the AW similar to that in the cortical projection of the retinal image. The retinal image in the AW is used as an input to the module for primary feature detection, which performs a function similar to the primary visual cortex. This module contains a set of neurons with orientationally selective receptive fields (ORFs) tuned to different orientations of the local edge. Neurons with the ORF, centered at the same point but with different orientation tuning, interact competi-

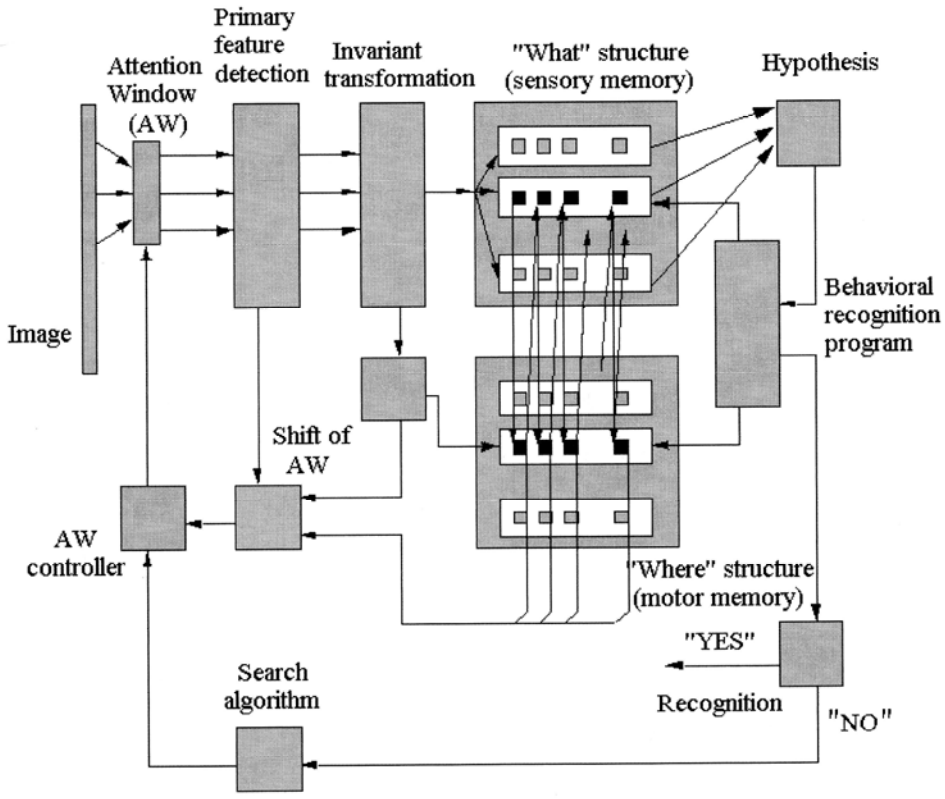

FIGURE 108.1 Schematic of the model.

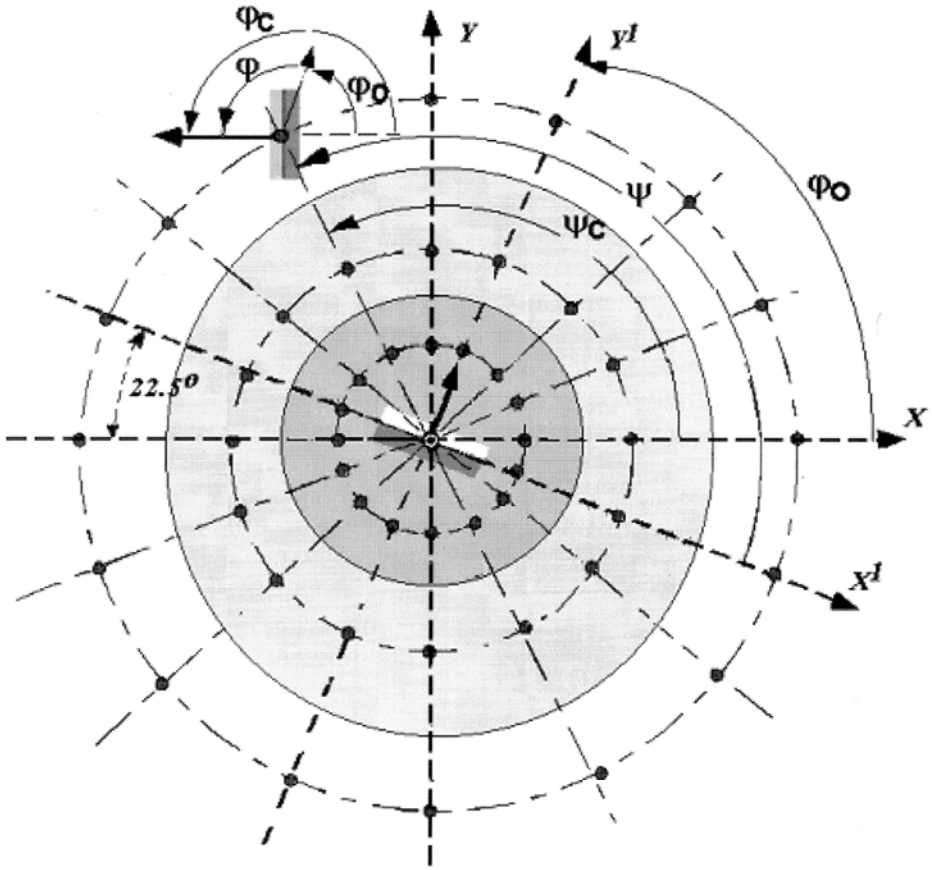

FIGURE 108.2 Schematic of the attention window (AW). The AW contains three areas with different resolution, which are separated by shadings: the central circle—fovea (has maximal resolution), the surrounding ring (shaded) with a lower resolution, and rest area with the lowest resolution. The context points are located at the intersections of 16 radiating lines and three concentric circles, each in a different resolution area. *X0Y* is the absolute coordinate system. The relative coordinate system X^1OY^1 (feature-based frame of reference) is attached to the basic edge in the center of the AW. The absolute parameters of one context edge, φ_c and ψ_c, and its relative parameters, φ and ψ, are shown. φ_o is the angular difference between the absolute and relative coordinate systems, which corresponds to the orientation of the basic edge in the absolute coordinate system.

tively due to strong reciprocal inhibitory interconnections. The orientation tuning of the winning neuron encodes the edge orientation at this point. The module for primary feature detection extracts a spatial pattern of edges, including the basic edge in the center of the AW and several context edges). This pattern of edges characterizes the retinal image at given point of fixation (Fig. 108.3c).

The modules just described form a low-level subsystem of the model. The next module performs an intermediate-level processing and completes the preattentive stage in the model. It transforms the set of primary features into invariant second-order features using a coordinate system (frame of reference) attached to the basic edge in the center of the AW and oriented along the brightness gradient of the basic edge (Fig. 108.2). The relative orientations and relative angular locations of the context edges with respect to the basic edge are considered as invariant second-order features.

The performance of the high-level subsystem and the entire model may be considered in three different modes: memorizing, search, and recognition.

In the memorizing mode, the image is processed at sequentially selected fixation points. At each fixation point, the set of edges is extracted from the AW, transformed into the invariant second-order features and stored in the sensory memory ("what" structure; see Fig. 108.1). The next position of the AW (next fixation point) is selected from the set of context points (Fig. 108.3d) and is also represented with respect to the coordinate system attached to the basic edge

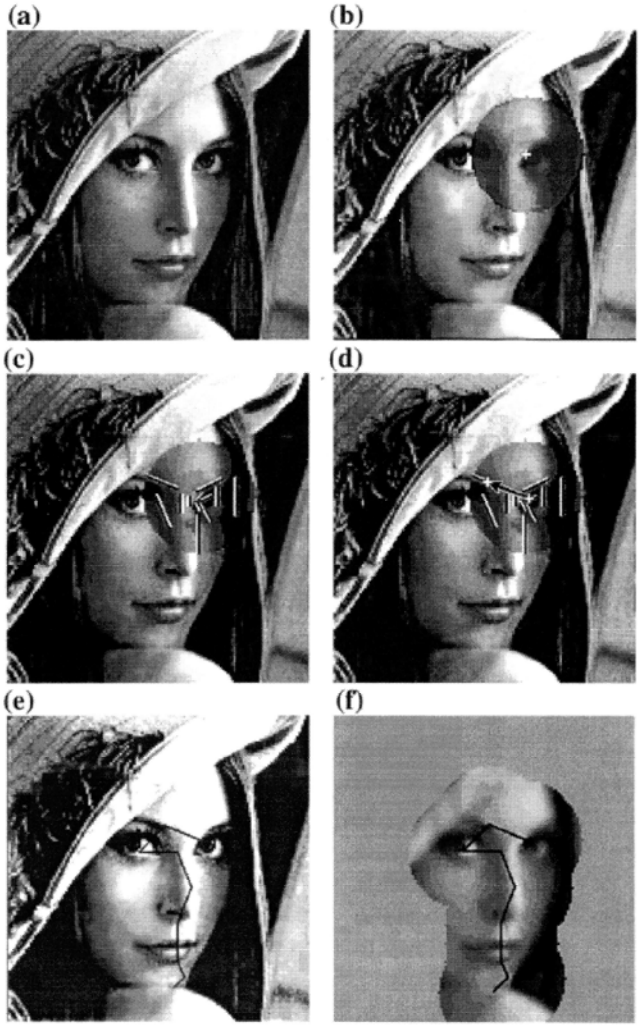

FIGURE 108.3 Processing of the image. (a) The initial image. (b) Image transformation within the AW (the retinal image) in one fixation point. (c) The basic and context edges detected from the retinal image. (d) The shift to the next fixation point (shown by the black arrow). (e) The scanpath of image viewing on the background of the initial image. (f) The same scanpath on the background of the sequence of retinal images along the scanpath.

(Fig. 108.2). A special module shifts the AW to the new fixation point via the AW controller playing the role of the oculomotor system (see Fig. 108.1). The scanpath of image viewing is formed as a series of sequential shifts of the AW (series of succads, see Fig. 108.3e–f). Each sequential shift of the AW (saccadic eye movement) is stored in the motor memory ("where" structure). As a result of the memorizing mode, the sequence of retinal images is stored in the "what" structure (sensory memory), and the sequence of AW movements is stored in the "where" structure (motor memory). These two types of elementary "memory traces" alternate in a chain that is considered to be as a behavioral program of recognition for the particular image (see Fig. 108.1).

In the search mode, the image is scanned by the AW under the control of a search algorithm. At each fixation, the current retinal image from the AW is compared to all retinal images of all objects stored in the sensory memory. The scanning of the image continues until an input retinal image similar to one of the stored retinal images is found at some fixation point. When such retinal image is found, a hypothesis about the image is formed and the model turns to the recognition mode.

In the recognition mode, the behavioral program is executed under the top-down control of attention. The recognition occurs by way of consecutive shifts of the AW (performed by the AW controller receiving information from the motor memory) and consecutive verification of the predicted retinal images recalled from the sensory memory. The scanpath of viewing in the recognition mode reproduces sequentially the scanpath of viewing in the memorizing mode. If a series of successful matches occurs, the object is considered recognized; otherwise, the model returns to the search mode.

B. Model Testing: Recognition of Faces and Objects in Complex Scenes

Gray-scale images of scene objects and of faces were used to test the ability of the model to recognize complex objects. Examples of these images are shown in Figs. 108.4a and 108.5a. In the memorizing mode, all tested images were sequentially presented to the model for memorizing. In the memorizing mode, the model viewed each image by way of sequential selection of fixation points and processing of the parts of the image within the AW centered at these points. Then, each memorized image was transformed (shifted, resized, and rotated) and presented to the model for recognition (see examples in Figs. 108.4b and 108.5b). Under these ideal conditions (without significant perturbations of the images or big occlusions), the model always recognized the image (see Figs. 108.4b and 108.5b). Only significant occlusions or perturbations that broke the scanpath (not allowing the required number of sequential successful matches) could cause an error when the model did not recognize the previously memorized image. These experiments have shown that the model is able to recognize practically any complex gray-scale image invariantly with respect to shift, 2D rotation, and scale. The scanpaths of image viewing in the memorizing mode are shown in Figs. 108.4a and 108.5a. In the recognition mode, the model executed the corresponding behav-

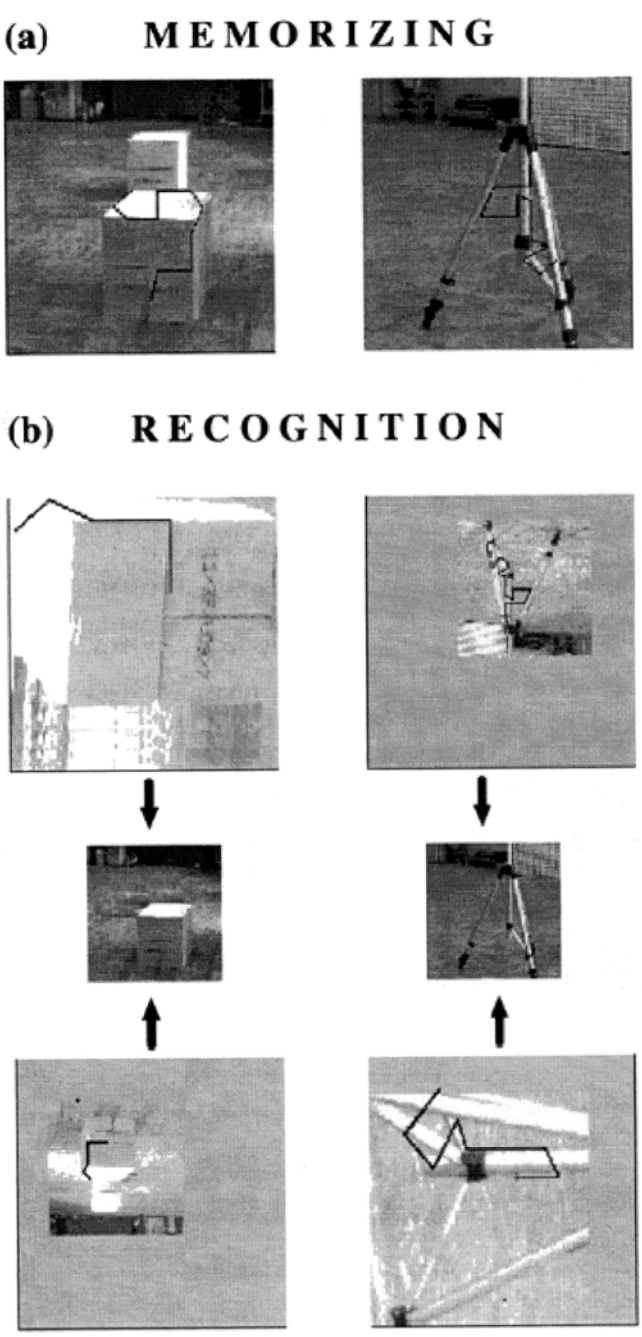

FIGURE 108.4 Examples of recognition of test images invariantly with respect to shift, rotation, and scale (a) The test images with the scanpaths in the memorizing mode. (b) The images presented for recognition (with the scanpaths during the recognition mode). These images were obtained from the test images by shifting, rotation, and scaling. The results of recognition are shown by arrays.

ioral program associated with the accepted hypothesis about the object. The scanpaths of viewing during the recognition mode were topologically similar to the scanpaths of viewing during the memorization of the same images (see Figs. 108.4b and 108.5b).

III. DISCUSSION: INVARIANT IMAGE REPRESENTATION AND RECOGNITION

The exciting ability of natural visual systems of invariant recognition has attracted the attention of scientists for more than 100 years. For several decades, the property of invariant recognition has been one of the major objectives and test criteria in different scientific areas from artificial neural networks to computer and robot vision. However, from a behavioral point of view, the ability of invariant recognition is not an ultimate goal but rather a tool that helps the system to plan and execute actions. This problem-oriented (task-driven) behavior requires a problem-oriented perception and recognition. According to Didday and Arbib (1975), the goal is "not to build a little internal copy of the visual scene, but rather to assist making the decision of what action (if any) the organism should next initiate." From this point of view, the absolute invariant recognition; in which exactly the same internal representation is achieved for an object irrespective of its size, location, orientation in space, and so forth, is useless. In that case, the organism will learn that the object is present in the scene, but will not know how to manipulate it. For example, the system will know that a cup is present, but will not know how to take it (where the handle is) and whether it is possible to pour some tea into it (how the cup is oriented in space). Thus, recognition should be considered as a process (behavior) during which the system either actively manipulates a noninvariant object representation in memory (by transforming it to match the external image view) or manipulates the external image using active eye and head movements. The resultant manipulations, used to fit the model to the object, give the system additional information about object location, orientation, size, and so forth.

On the other hand, a lack of invariant representations makes the task of recognition practically unresolved. How would the system know which one of the majority of noninvariant models stored in memory to take for manipulations in order to match the object and how long to manipulate the selected model before making the decision to take another model? The natural visual system evidently solves this dilemma by way of some "smart" combination of these two opposite approaches.

A possible (and hopefully plausible) way for such a combination may be based on the invariant representation of object's elements in the vicinity of fixation points (within the AW) and on the use of object manipulations (eye movements and shifts of the AW) in order to represent spatial relationships. The latter idea was used in the reviewed model. The model holds

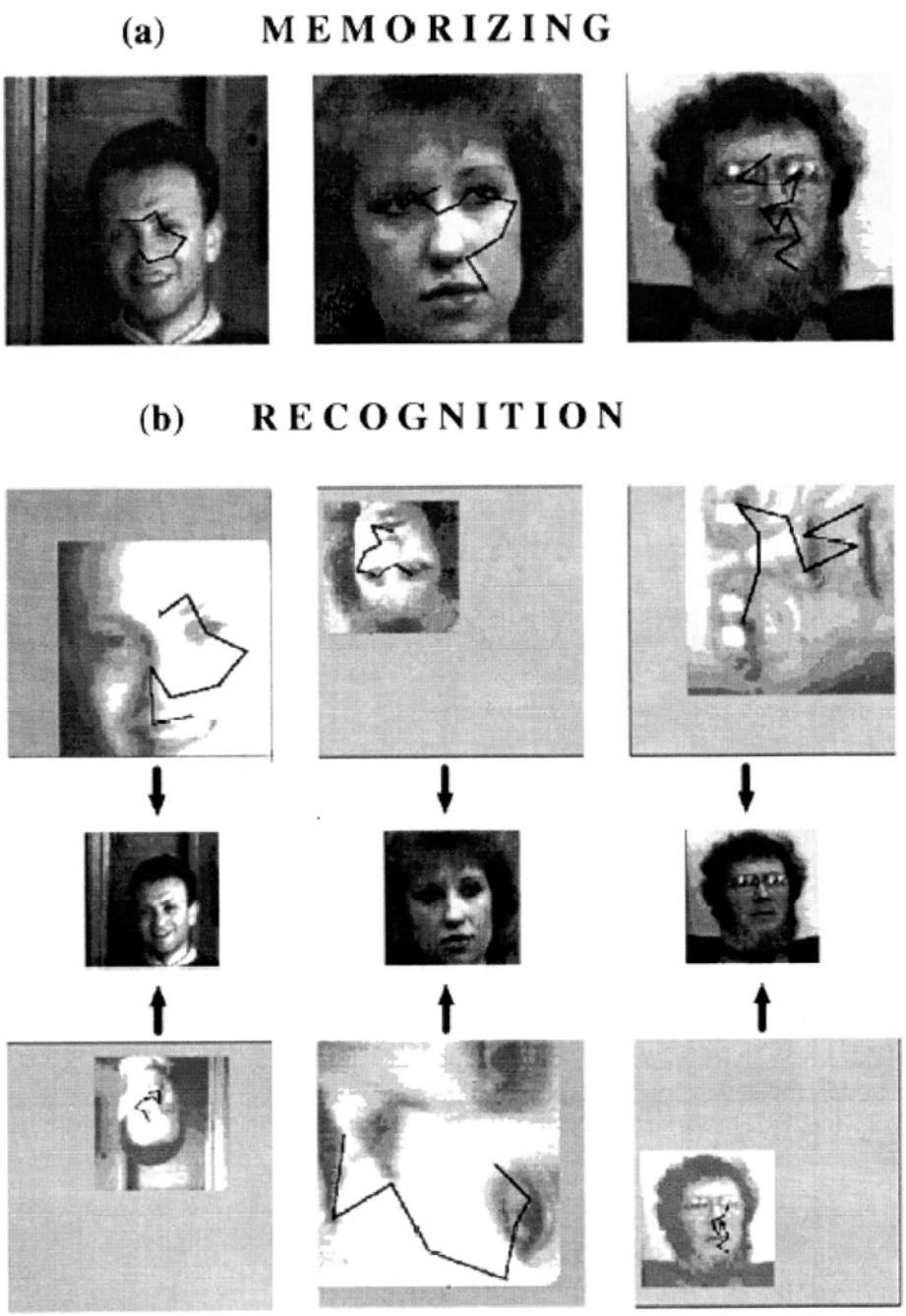

FIGURE 108.5 Examples of recognition of faces invariantly with respect to shift, rotation, and scale. (a) The test images with scanpaths in memorizing mode. (b) The images presented for recognition (with the scanpaths during the recognition mode).

invariantly represented image fragments in the sensory memory. Each fragment is associated with a certain object and with a certain action that the system should execute according to the behavioral recognition program that is associated with the object. The initial identification of a known fragment (invariantly represented in memory) gives a start of the behavioral recognition process, which is executed under the top-down control of attention. The recognition process provides information about object location and orientation in space. The system architecture used in the model, coupled with the behavioral algorithms of image memorizing and recognition and with the feature-based reference frame, allows the system both to recognize objects invariantly with respect to their position and orientation in space and to manipulate

objects in space using absolute parameters of the basic feature at the fixation point and relative spatial relationships recalled from the motor memory.

In the reviewed model, the algorithm for the invariant representation of the retinal image within the AW is based on the encoding of relative (with respect to the basic edge at the fixation point) orientations and angular positions of the detected edges. In other words, each basic edge at the fixation point is considered in the context of a set of other edges in the retinal image. With the decrease in resolution toward the retinal periphery, a more detailed and precise representation of image partition in the vicinity of the fixation point is considered in the context of a coarser generalized representation of a lager part of the image (or the entire image).

In order to memorize a particular object in the image or scene containing several objects and/or a complex background, the model should select only the points of fixation that belong to the same object. The current version of our model does not do this in general. In order to make this certain, in the memorizing mode the model should deal with images containing single objects with a uniform background. Then, in the recognition mode, the model should be able to recognize these objects in multiobject scenes with complex backgrounds. In contrast, the natural visual system uses special mechanisms that provide object separation independent of or even before object recognition (stereopsis and binocular depth perception; analysis of occlusions during head and body movements, color and texture analysis, etc.). Additional mechanisms that separate the objects in the image from one another and from the background should be incorporated in the model to allow memorizing objects in complex multiobject images. These mechanisms will prevent a selection of fixation points outside the object of interest.

In conclusion, the reviewed model provides important insights into the role of behavioral aspects for invariant pattern recognition. The basic algorithmic ideas of the model and approach used may be applied to computer and robot vision systems aimed toward invariant image recognition.

References

Burt, P. J. (1988). Smart sensing within a pyramid vision machine. *Proc. IEEE* **76**, 1006–1015.

Didday, R. L., and Arbib, M. A. (1975). Eye movements and visual perception: A two visual system model. *Int. J. Man-Machine Studi.* **7**, 547–569.

Hinton, G. E., and Lang, K. J. (1985). Shape recognition and illusory conjunctions. *In* "Proceedings of the 9th International Joint Conference on Artificial Intelligence" (Aravind K. Joshi Ed.), pp. 252–259. Morgan Kaufmann, Los Angeles, CA.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular integration and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154.

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." W. H. Freeman, New York.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends Neurosci.* **6**, 414–417.

Neisser, V. (1967). "Cognitive Psychology." Appleton, New York.

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

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. *In* "Human and Machine Vision" (J. Beck, B. Hope, and A. Rosenfeld, Eds.), pp. 545–567. New York: Academic Press.

Posner, M. I., and Presti, D. E. (1987). Selective attention and cognitive control. *Trends Neurosci.* **10**, 13–17.

Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., and Shevtsova, N. A. (1998). A model of attention-guided visual perception and recognition. *Vis. Res.* **38**, 2387–2400.

Treisman, A. M., and Gelade, G. (1980). A feature integration theory of attention. *Cogr. Psychol.* **12**, 97–136.

Ungerleider, L. G., and Mishkin, M. (1982). Two cortical visual systems. *In* "Analysis of Visual Behavior" (D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds.), pp. 549–586. MIT Press, Cambridge, MA.

Yarbus, A. L. (1967). "Eye Movements and Vision." Plenum, New York.