PERGAMON

Vision Research 38 (1998) 2387–2400

Vision Research

A model of attention-guided visual perception and recognition

I.A. Rybak [a,*], V.I. Gusakova [b], A.V. Golovan [c], L.N. Podladchikova [b], N.A. Shevtsova [d]

[a] *E. I. du Pont de Nemours and Co., Central Research Department, Experimental Station E-328/B31, Wilmington, DE 19880-0328, USA*
[b] *A. B. Kogan Research Institute for Neurocybernetics, Rostov State University, 194/1 Stachka Ave., Rostov-on-Don 344090, Russia*
[c] *MBTI, George Mason University, Fairfax, VA 22030, USA*
[d] *Institute of Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA*

Received 9 January 1997; received in revised form 25 June 1997; accepted 16 December 1997

Abstract

A model of visual perception and recognition is described. The model contains: (i) a low-level subsystem which performs both a fovea-like transformation and detection of primary features (edges), and (ii) a high-level subsystem which includes separated 'what' (sensory memory) and 'where' (motor memory) structures. Image recognition occurs during the execution of a 'behavioral recognition program' formed during the primary viewing of the image. The recognition program contains both programmed attention window movements (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. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Visual perception; Invariant recognition; Attention; Eye movements; Scanpath

1. Introduction

1.1. Behavioral paradigm

It is known that the density of photoreceptors in the retina is greatest in the central area (fovea) and decreases to the retinal periphery whereas the size of neuronal receptive fields in the retinal output and in the cortical map of the retina increases to the periphery [1–4]. As a result, the resolution of the image representation in the visual cortex is highest for the part of the image projected onto the fovea and decreases rapidly with distance from the fovea center. During visual perception and recognition, human eyes move and successively fixate at the most informative parts of the image which therefore are processed with the highest resolution. At the same time, the mechanism of visual attention uses information extracted from the retinal periphery for selection of the next eye position and control of eye movement. Thus, the eyes actively perform problem-oriented selection and processing of information from the visible world under the control of visual attention [5–13]. Consequently, visual perception and recognition may be considered as behavioral processes and probably cannot be completely understood in the limited context of neural computations without taking into account the behavioral and cognitive aspects.

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 external) signals which are expected to arrive in response to each action. The brain uses these chains as 'behavioral programs' in subconscious 'behav-

[*] Corresponding author. Fax: +1 302 6958901; e-mail: rybaki@eplrx7.es.dupont.com.

0042-6989/98/$19.00 © 1998 Elsevier Science Ltd. All rights reserved.
PII: S0042-6989(98)00020-0

ioral recognition' when the object is (or is assumed) known. This 'behavioral recognition' has two basic stages: conscious selection of the appropriate behavioral program and the subconscious execution of the program. Matching the expected (predicted) sensory signals with the actual sensory signals, arriving after each motor action, is the essential procedure in the program execution.

The above behavioral paradigm was formulated and developed in the context of visual perception and recognition in a series of significant works [9,13–15]. Using Yarbus' approach [13], Noton and Stark [9] compared the individual scanpaths of human eye movements in two phases: during memorizing and 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 [9] and Didday and Arbib [14] 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 [16], Mishkin et al. [17], Van Essen [4], and Kosslyn et al. [15] presented neuroanatomical and psychological data complementary to the above behavioral concept. It was found that the higher levels of the visual system contain two major pathways for visual processing called '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 [4,15,17,16]. The behavioral concept joined with this neuro-anatomical theory provides: (i) the explicit functional coupling between the low-level vision (foveal structure of the retino-cortical projection, orientation selectivity in the visual cortex, etc.) and the high-level brain structures involved in visual perception and recognition; (ii) the clear functional role of visual attention in the coupling between the low- and high-levels of the visual system.

The behavioral concept of visual perception and recognition has been widely accepted in the field of robot vision ([5,18–23], and others). Rimey and Brown [23] presented a detailed analysis of the behavioral concept and developed a model of selective attention based on this concept (the Augmented Hidden Markov Model). The present model was developed at approximately the same time (in 1990–1991), when all authors worked together in A.B. Kogan Research Institute for Neurocybernetics at Rostov State University.

1.2. Image features, invariant representation and frames of reference

The question of what features are detected from 2D retinal images at the preattentive stage to represent the shape of 3D objects is still open. Beginning with the classic work of Hubel and Wiesel [24,25], 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 early theories based on neurophysiological data assumed that the visual system detects relatively simple features at the preattentive stage and uses some attention mechanisms of a serial type to bind the simple features into more complex shape features [7,8,12]. Alternatively, recent psychological studies have shown that early visual processes are much more sophisticated than previously assumed. Certain relations among features are detected preattentively including some 2D feature combinations reflecting elements of 3D shapes [26,27]. These data support the idea that at the end of the preattentive stage the retinal image is represented by the primary features in some mutual spatial relations or spatial compositions (patterns). The spatial patterns of edges may represent 2D projections of elementary 3D shapes (e.g. vertices) and more complex combinations of 3D angles. The detection of spatial patterns of edges at the preattentive stage may contribute to 3D scene perception in addition to other mechanisms (stereopsis and binocular depth perception; color and texture analysis; analysis of occlusions during head or body movements, etc.).

One key issue of visual recognition is the mechanism used for invariant image representation. Marr [28], Palmer [29], Hinton and Lang [30] 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 shift, making it difficult to recognize the object. As a result, previous models of recognition using the object-based frame of reference demonstrated invariant recognition of only simple objects (letters, binary objects without background, etc.) to which such a frame of reference is easily attached (for example, see Refs. [31–34]).

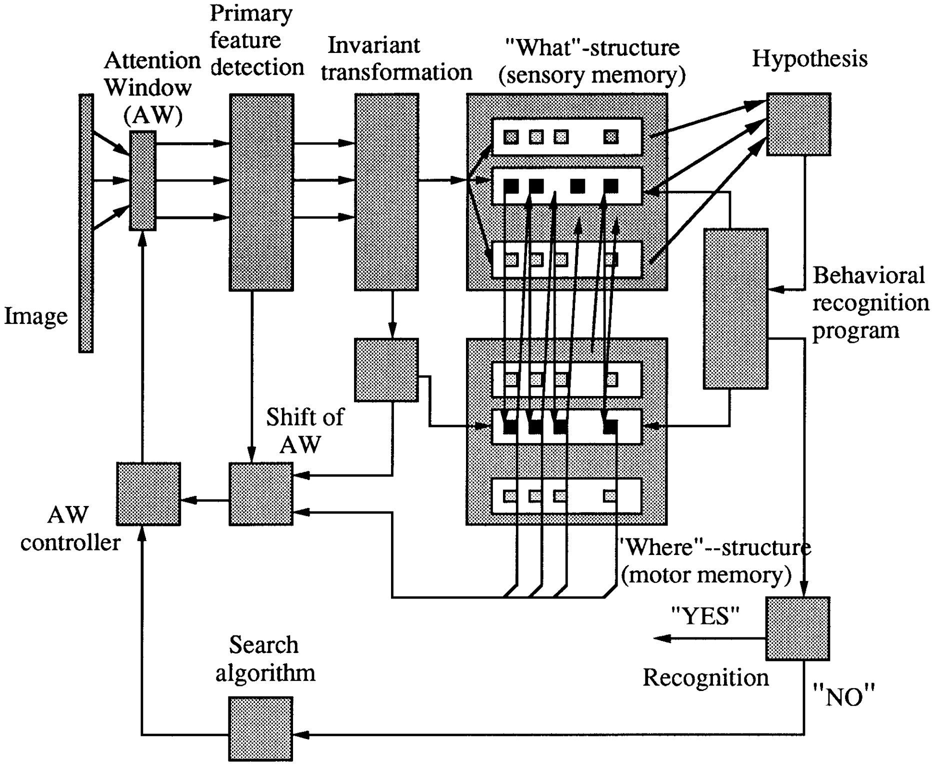
Primary feature detection
Invariant transformation
"What"-structure (sensory memory)
Hypothesis
Attention Window (AW)
Image
Behavioral recognition program
Shift of AW
AW controller
"Where"--structure (motor memory)
"YES"
Recognition
"NO"
Search algorithm

Fig. 1. Schematic of the model.

In our model, a set (spatial composition or pattern) of edges is extracted from the retinal image at each fixation. Image representation at the fixation point is based on the assumption that the features (edges) extracted from the retinal periphery have 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, for example, see Ref. [5]). The other function of these edges, unique to the present model, is to provide a context ('context features') for the 'basic' feature (edge) in the center of the fovea.

The relative orientations of context edges and their relative angular locations with respect to the basic edge are invariant to rotation and size, and may be used as second-order invariant features of the image at the current point of fixation. Thus, instead of the object-based frame of reference, we use a feature-based frame of reference attached to the basic edge at the fixation point. Since 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. Moreover, the feature-based frame of reference coupled with the multiplicity of fixation points along the scanpath allows image recognition from a part of the image (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. Thus, the stability of recognition increases with the number of fixations.

2. Model

2.1. General description of the model

A functional diagram of the model is shown in Fig. 1. The attention window[1] (AW) performs a primary transformation of the image into a 'retinal image' at the fixation point. The primary transformation provides a decrease in resolution of the retinal image from the center to the periphery of the AW, that simulates the decrease in resolution from the fovea to the retinal periphery in the cortical map of the retina. 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 (ORF) tuned to different orientations of the local edge. Neurons with the ORF, centered at the same point but with different orientation tuning, interact competitively due to strong reciprocal inhibitory interconnections. The orientation tuning of the 'win-

[1] The present model focuses on the attention mechanisms associated with eye movements and foveation. Covert attention, which is not associated with eye movements and the structure of the retina, has not been in the scope of our consideration although a number of ideas and algorithms used may relate to covert attention as well. The term 'attention window' used in our model directly relates to the part of the image projected onto the retina and differs from the same term used in publications on covert attention.

ning neuron' encodes the edge orientation at this point. The module for primary feature detection extracts a set of edges (one 'basic' edge in the AW center and several 'context' edges) which characterizes the retinal image at given point of fixation.

The modules described above 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. 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). The next position of the AW (next fixation point) is selected from the set of context points and is also represented with respect to the coordinate system attached to the basic edge. A special module shifts the AW to a new fixation point via the AW controller playing the role of the oculomotor system. Each relative shift of the AW ('eye movement') is stored in the motor memory ('where'-structure). As a result of the memorizing mode, the whole 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 which is considered as a 'behavioral recognition program' for the memorized image.

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 a 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 by way of consecutive shifts of the AW (performed by the AW controller receiving information from the motor memory) and consecutive verification of the expected 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 recognized, otherwise the model returns to the search mode.

2.2. Primary transformation: formation of the retinal image within the AW

The retinal image results from the initial image $I=\{x_{ij}\}$ by way of a special transformation used to obtain a decrease in resolution from the AW center to its periphery. To represent a part D of the image $((i,j)\in D)$ at resolution level $l(l\in\{1, 2, 3, 4, 5\})$ we used the recursive computation of the Gaussian-like convolution [5] at each point of D:

$$\begin{aligned}
x_{ij}^{(1)} &= \quad x_{ij}\\
x_{ij}^{(2)} &= \sum_{p=-2}^{p=2}\sum_{q=-2}^{q=2} g_{pq}\cdot x^{(1)}_{i-p,j-q}\\
&\ \ \vdots\\
x_{ij}^{(l)} &= \sum_{p=-2}^{p=2}\sum_{q=-2}^{q=2} g_{pq}\cdot x^{(l-1)}_{i-p\cdot 2^{l-1},j-q\cdot 2^{l-1}}
\end{aligned}\tag{1}$$

where the coefficients of convolution belong to the following matrix [5]:

$$[g_{pq}] = \begin{bmatrix} 1 & 4 & 6 & 4 & 1\\ 4 & 16 & 24 & 16 & 4\\ 6 & 24 & 36 & 24 & 6\\ 4 & 16 & 24 & 16 & 4\\ 1 & 4 & 6 & 4 & 1\end{bmatrix}\cdot\frac{1}{256}$$

$$(p \text{ and } q = -2, -1, 0, 1, 2). \tag{2}$$

In the model, the primary image transformation maps the initial image $I=\{x_{ij}\}$ into the retinal image $I^{\mathrm{R}}(n)=\{x^{\mathrm{R}}_{ij}(n)\}$ at each nth fixation point. The position of the fixation point $(i_{\mathrm{o}}(n), j_{\mathrm{o}}(n))$ and the resolution level $l_{\mathrm{o}}(n)$ in the vicinity of that point are considered to be parameters of the retinal image. The central point $(i_{\mathrm{o}}(n), j_{\mathrm{o}}(n))$ is surrounded by three concentric circles whose radii are functions of $l_{\mathrm{o}}(n)$:

$$\begin{aligned}
R_0(l_{\mathrm{o}}) &= 1.5\cdot 2^{l_{\mathrm{o}}},\\
R_1(l_{\mathrm{o}}) &= 1.5\cdot 2^{l_{\mathrm{o}}+1},\\
R_2(l_{\mathrm{o}}) &= 1.5\cdot 2^{l_{\mathrm{o}}+2}.
\end{aligned}\tag{3}$$

The retinal image at the nth fixation point $I^{\mathrm{R}}(n)=\{x^{\mathrm{R}}_{ij}(n)\}$ is formed from $I=\{x_{ij}\}$ as follows:

$$x^{\mathrm{R}}_{ij}(n) = \begin{cases} x_{ij}^{l_{\mathrm{o}}(n)}, & \text{if } \rho_{ij}(n)\le R_0(l_{\mathrm{o}})\\ x_{ij}^{l_{\mathrm{o}}(n)+1}, & \text{if } R_0(l_{\mathrm{o}})<\rho_{ij}(n)\le R_1(l_{\mathrm{o}})\\ x_{ij}^{l_{\mathrm{o}}(n)+2}, & \text{if } R_1(l_{\mathrm{o}})<\rho_{ij}(n)\le R_2(l_{\mathrm{o}})\end{cases}\tag{4}$$

where

$$\rho_{ij}(n) = \sqrt{(i - i_o(n))^2 + (j - j_o(n))^2}. \qquad (5)$$

Thus, the initial image is represented in the AW: with the highest resolution $l = l_o(n)$ within the central circle ('fovea'), with lower resolution $l = l_o(n) + 1$ within the first ring surrounding the central circle, and with the lowest resolution $l = l_o(n) + 2$ within the second ring.

The areas of different resolution are shown in Fig. 2, where the central circle -'fovea' and the first ring are separated by different shading patterns. Fig. 3(b) shows an example of the retinal image at one point of fixation.

2.3. Detecting primary features

The module for primary feature detection performs a function similar to the function of the primary visual cortex containing orientationally selective neurons. According to the Hubel and Wiesel theory, neurons tuned to the same orientation are packed into orientation columns which, in turn, comprise the retinotopically organized processing module of the visual cortex-hypercolumn [24,25]. Therefore, visual cortex neurons detect the orientation of edges at each point of the retinal image. In our model, edges are detected by a network of orientationally selective neurons and are considered to be the primary features of the image. Each edge is detected with resolution dependent on the position of the edge in the retinal image.

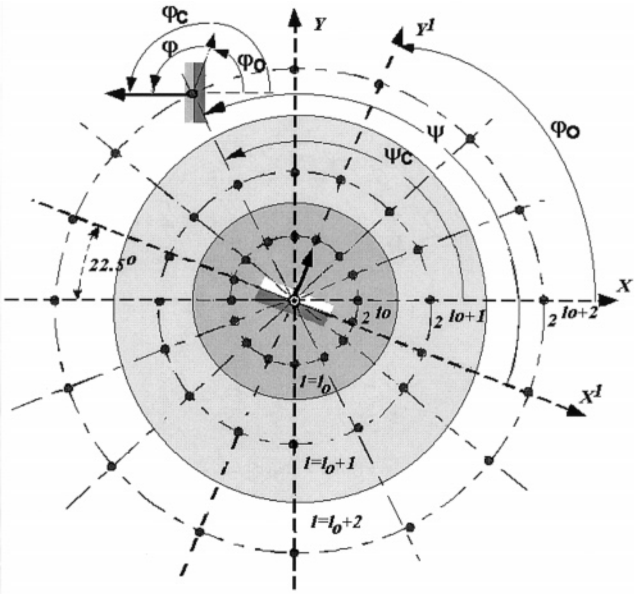

Fig. 2. Schematic of the attention window (AW). The areas of different resolution (central circle -'fovea' and first ring) are separated by shadings. The context points are located at the intersections of sixteen radiating lines and three concentric circles, each in a different resolution area. $X0Y$ is the absolute coordinate system. The relative coordinate system X^1OY^1 ('feature-based frame of reference') is attached to the basic edge at 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.

The orientationally selective receptive field (ORF) of the neuron with coordinates (i, j) tuned to the orientation α ($\alpha = 0, 1, 2, \ldots, 15$) is formally described in our model using the algorithm by Grossberg, Mingolla and Todorovic [35]. The discrete angle step of 22.5° is considered as a unit in all angle measurements. The ORF is described as a difference between two Gaussian convolutions with spatially shifted centers. The input signal to the neuron tuned to the orientation α is:

$$Y_{ij\alpha} = \sum_{pq} x^R_{pq} \cdot (G^+_{pqij\alpha} - G^-_{pqij\alpha}), \qquad (6)$$

where

$$G^+_{pgij\alpha} = \exp(-\gamma^2 \cdot ((p - i - m_\alpha)^2 + (q - j - n_\alpha)^2));$$

$$G^-_{pgij\alpha} = \exp(-\gamma^2 \cdot ((p - i + m_\alpha)^2 + (q - j + n_\alpha)^2)). \qquad (7)$$

In Eq. (7), γ is a reciprocal variance. The parameters m_α and n_α depend on the ORF orientation α:

$$m_\alpha = \mathrm{d}(l) \cdot \cos(2 \cdot \pi \cdot \alpha/16);$$

$$n_\alpha = \mathrm{d}(l) \cdot \sin(2 \cdot \pi \cdot \alpha/16), \qquad (8)$$

where d(l) defines the distance between the center of each Gaussian and the center of the ORF and depends on the resolution level l in a given area of the retinal image:

$$\mathrm{d}(l) = \max\{2^{l-2}, 1\}. \qquad (9)$$

Sixteen neurons, whose ORF have the same location but different orientations, interact competitively due to strong reciprocal inhibitory connections:

$$\tau \cdot \frac{\mathrm{d}}{\mathrm{d}t} V_{ij\alpha} = -V_{ij\alpha} + Y_{ij\alpha} - b \cdot \sum_{\substack{k=0 \\ k \neq \alpha}}^{15} Z_{ijk} - T;$$

$$Z_{ij\alpha} = f(V_{ij\alpha}); \quad \alpha = 0, 1, 2, \ldots, 15, \qquad (10)$$

where $V_{ij\alpha}$ and $Z_{ij\alpha}$ are the membrane potential and output of the neuron (i, j) with the ORF tuned to the orientation α, respectively; b is a coefficient characterizing the reciprocal inhibition ($b > 1$); T is the neuron threshold; τ is the time constant; $f(V) = V$ if $V \geq 0$, otherwise $f(V) = 0$.

The possible steady state solutions of system (10) are: (i) all $\bar{Z}_{ij\alpha} = 0$ (if all $Y_{ij\alpha} < \mathrm{T}$) or (ii) only one $\bar{Z}_{ij\alpha} = Y_{ij\alpha} - T > 0$ at $\alpha = \varphi^*$ for which $Y_{ij\alpha}$ is maximal ($\bar{Z}_{ij\alpha} = 0$, if $\alpha \neq \varphi^*$). In the first case, there are no edges at the point (i, j); in the second case, there is an edge with the orientation φ^* at this point.

At each nth fixation (AW position) oriented edges are detected: at the fixation point $(i_o(n), j_o(n))$ (the 'basic edge' in the center of the AW) and at 48 'context'

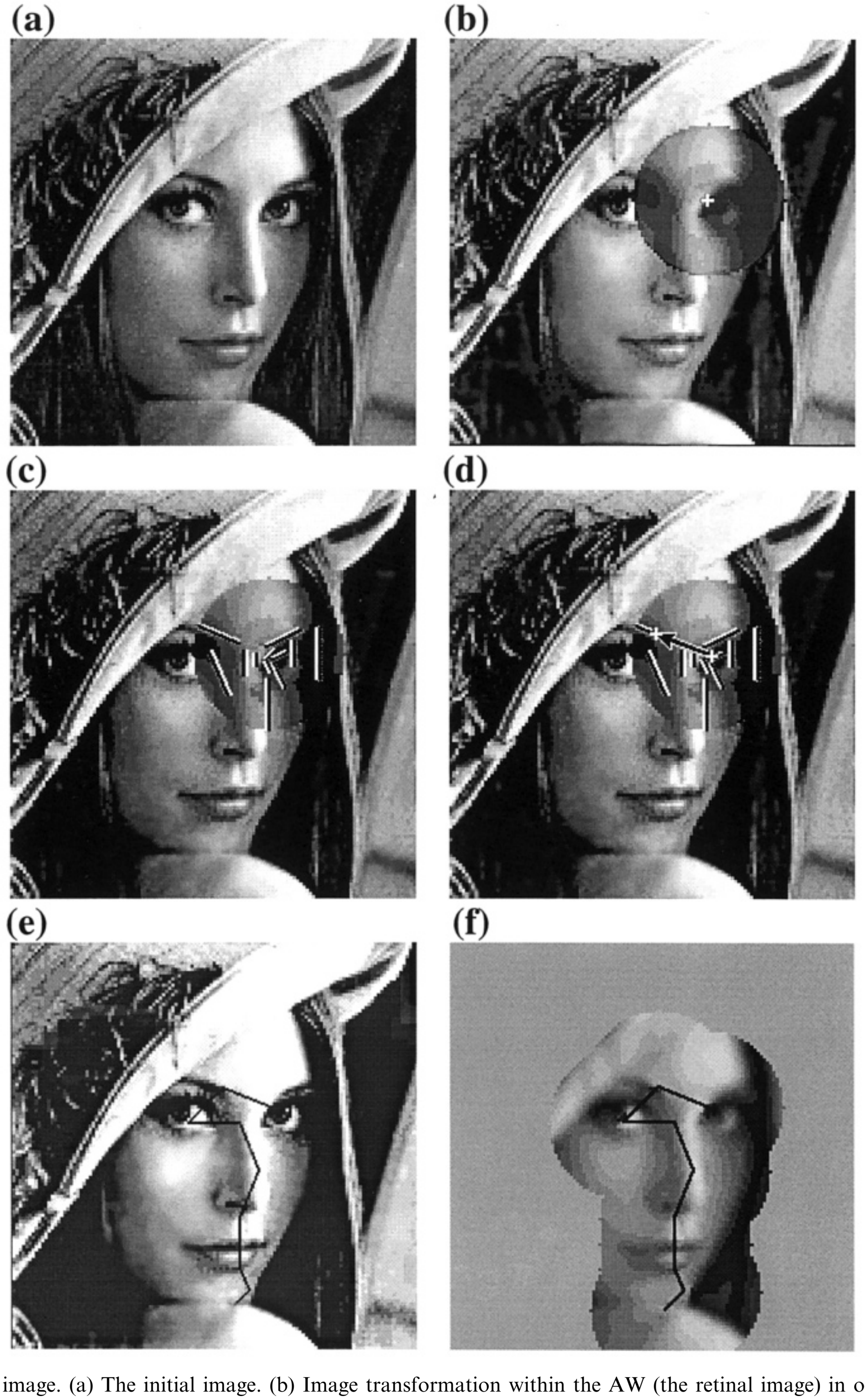

Fig. 3. Processing of the image. (a) The initial image. (b) Image transformation within the AW (the retinal image) in one fixation point. (c) The basic and context edges detected from the retinal image. (d) The shift to the next fixation point (shown by the black arrow). (e) The scanpath of image viewing on the background of the initial image. (f) The same scanpath on the background of the sequence of retinal images along the scanpath.

points which are located at the intersections of sixteen radiating lines (with the angle step of 22.5°) and three concentric circles, each in a different resolution area (see Fig. 2). The radii of these circles (R_{00}, R_{01} and R_{02}) exponentially increase:

$$R_{00}(l_o) = 2^{l_o};$$

$$R_{01}(l_o) = 2^{l_o+1};$$

$$R_{02}(l_o) = 2^{l_o+2}. \qquad (11)$$

Context edges located in the smallest circle are detected with the same resolution as the basic edge; the resolution with which the other edges are detected is determined by their positions in the retinal image. The

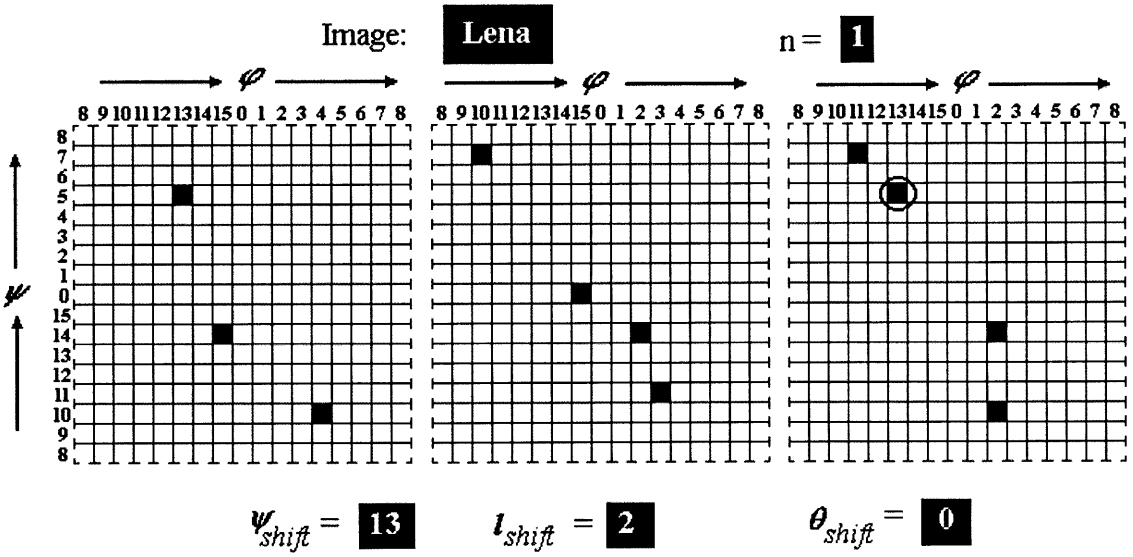

Fig. 4. Invariant representation of a set of edges on the rectangular maps of surfaces of three tori. Relative orientation φ and relative angular location ψ change in horizontal and vertical directions, respectively. The points on each map correspond to the edges extracted from one area of resolution (one circle in Fig. 2). The arrow of points on the maps correspond to the set of edges in Fig. 3(b). The next fixation point (the same as in Fig. 3(d)) is indicated by circle. The parameters of the AW shift to the next fixation point are shown at the bottom.

basic and context edges in one fixation point (the same as in Fig. 3(b)) are shown in Fig. 3(c) by doubled white and black segments whose lengths increase to the AW periphery with the decrease of resolution.

2.4. Invariant representation and comparison of the retinal images: a feature-based frame of reference

Let us attach the absolute coordinate system XOY to the AW center (Fig. 2). The basic edge in the AW center may be represented by the pair of parameters (φ_o, l_o), where: φ_o ($\varphi_o \in \{0, 1,\ldots,15\}$) is the orientation of the basic edge (represented by the angle between the axis $\vec{OX}$ and the vector of brightness gradient of the basic edge, see Fig. 2) and l_o ($l_o \in \{1, 2, 3\}$) is the level of resolution in the central area of the AW. Each context edge can be represented in the absolute coordinate system by the three parameters: (φ_c, ψ_c, l) where: φ_c is the orientation of the context edge; ψ_c characterizes the angular location of the edge in XOY (see Fig. 2); l is the level of resolution in the area of the edge; φ_c and $\psi_c \in \{0, 1, \ldots, 15\}$.

Let us now attach the 'relative' coordinate system X^1OY^1 so that the axis $\vec{OY}^1$ is directed along the vector of brightness gradient of the basic edge ('feature-based frame of reference'; see Fig. 2). The context edge (φ_c, ψ_c, l) may be represented with respect to the relative coordinate system by the parameters (φ, ψ, λ) (see Fig. 2), where: φ is the relative orientation of the context edge; ψ is its relative angular location; λ characterizes the relative distance from the AW center. These parameters are calculated as follows:

$$\begin{aligned}
&\varphi = \mathrm{mod}_{16}(\varphi_c - \varphi_o + 16);\\
&\psi = \mathrm{mod}_{16}(\psi_c - \varphi_o + 20);\\
&\lambda = l - l_o;\\
&\varphi,\ \psi \in \{0, 1, \ldots, 15\};\\
&\lambda \in \{0, 1, 2\}. \qquad (12)
\end{aligned}$$

Thus, the retinal image within the AW at nth fixation point can be invariantly represented by three arrays of pairs of numbers $\{\varphi_k(n), \psi_k(n)\}_\lambda$; $k = 0, 1,\ldots, 15$; $\lambda = 0, 1, 2$.

Since both $\varphi_k(n)$ and $\psi_k(n)$ change periodically, each nth retinal image may be invariantly represented by a set of points on the surfaces of three tori (each torus for one magnitude of λ). An example of such a representation for the set of edges in Fig. 3(c) is shown in Fig. 4 by the arrays of points on the rectangular maps of three toroidal surfaces. The 'toroidal' method was used in a complex version of the model for invariant representation of retinal images in the high-level subsystem (sensory memory). The patterns of points on the toroidal surfaces represent retinal images in an invariant form. Most classical neural network classifiers (NNC) may be taught to recognize such patterns, and be used in the high-level subsystem of the model for comparison of the arriving retinal images with those stored in the sensory memory. The chosen NNC should be able to recognize/compare patterns with disturbances and noise (that most classical NNC types can do well; see Ref. [36]). However, in this case there is no necessity in pattern recognition invariant to shift, rotation and scale (that still is an unresolved problem in the classical

NNC). In a more complex version of our model, the classical Hopfield neural network [37] was incorporated into the high-level subsystem of the model and successfully performed recognition (comparison) of sequentially arriving retinal images during both modes: search and recognition.

In a simplified version of the model, each nth retinal image is invariantly represented by three 16 component vectors $\vec{\varphi}_{\psi}(n)$ whose components $\varphi_k(n)$ are arranged with the increase of $\psi_k(n)$:

$$\vec{\varphi}_{\psi\lambda}(n) = \{\varphi_0(n);\ \varphi_1(n);\ \ldots;\ \varphi_{16}(n)\}_{\lambda};\quad \lambda = 1, 2, 3. \tag{13}$$

These vectors are stored in the 'what' structure (sensory memory). The following function has been used to evaluate a difference between the stored retinal image represented in the sensory memory by the vectors, $\vec{\varphi}_{\psi\lambda} = \{\varphi_0; \varphi_1; \ldots; \varphi_{16}\}_{\lambda}$, and the arriving retinal images represented by the vectors $\vec{\varphi}^{\,*}_{\psi\lambda} = \{\varphi^*_0; \varphi^*_1; \ldots; \varphi^*_{16}\}_{\lambda}$:

$$F\left(\vec{\varphi}_{\phi\lambda}, \vec{\varphi}^{\,*}_{\psi\lambda}\right) = \frac{1}{N}\cdot\sum_{\lambda}\sum_{\psi}\frac{1}{1 + 8\cdot\sin^2\left(\frac{\pi}{16}\left(\vec{\varphi}_{\phi\lambda} - \vec{\varphi}^{\,*}_{\psi\lambda}\right)\right)}. \tag{14}$$

Only edges (vector components) present in the stored retinal image are taken into account in Eq. (14); N is the number of these edges. The compared retinal images are considered similar if the value of F exceeds a threshold of recognition Thr found experimentally.

2.5. Shifting the AW: selection and representation of the next fixation point

In the memorizing mode, the model selects each next fixation point from the set of context points in the current retinal image (Fig. 2). Thus, the current context points are considered to be potential targets for the next shift of the AW.

It is necessary to note that the selection of the next fixation point from potential targets in the memorizing mode (when the system 'sees' an image for the first time) relates to very complicated, fundamental problems of visual search. Wolfe and Cave [38–40] argued that the attention mechanism for this selection uses both bottom-up information from the image and top-down information from the high-level visual structures. In Guided Search [38], they also made an attempt to formalize a top-down selection mechanism under conditions when the target features of the object have been specified. However, in general the top-down control of attention relates to the high-level cognitive processes which are poorly understood and weakly formalized. At the same time, it is clear that the selection of fixation points in memorizing mode should be partly predefined by the abstract knowledge of objects and the task at hand (problem-oriented selection). The present model has been directed toward modeling of the behavioral, structural and functional aspects of visual recognition with a special focus on possible mechanisms for invariant recognition. While developing the model, we did not intend to implicitly concern the complex semantic problems of visual search. But, because of the deep relationships between visual search and recognition, we would not be able to build a working model without incorporating a simplified visual search mechanism. However, the key problems of visual search were beyond the scope of the present model.

The present model does not have the top-down control of attention. At each fixation during memorizing mode, all context points attract the AW, each to its own position. The attractive effects of context points on the AW reciprocally inhibit each other, and the context point with the strongest 'attracting force' wins 'the competition for the AW' and attracts it to its position. Attracting force of the kth context point, A_k, is defined as follows:

$$A_k = a_1\cdot\frac{Z_k}{Z_{\max}} + a_2\cdot\frac{\lambda_k}{2} + a_3\cdot\eta_{ij}(n) + a_4\cdot\chi_{ij}. \tag{15}$$

The first two terms in Eq. (15) provide the dependence of the attracting force on the content of the image (bottom-up control of attention): the first term determines the normalized value of brightness gradient in the context point, which, in turn, is defined by the output of the corresponding neuron-winner in the module for primary feature detection; the second term determines the relative distance of the context edge from the center of the retinal image.

The third term was incorporated in Eq. (15) to prevent 'cycling'. The function $\eta_{ij}(n)$ determines a 'novelty' of the vicinity of the context point. After shifting the AW to the next fixation point, the function $\eta_{ij}(n)$ for all points within a vicinity of the previous fixation point falls to zero and then slowly recovers with time to the value of one. This function provides the 'inhibition of return' effect which has been demonstrated in a number of experiments starting from Posner and Cohen [41].

The function χ_{ij} in the fourth term ($\chi_{ij}\in[0, 1]$) predefines a 'semantic significance' of the area which includes the context point. This function is defined in advance for each image if $a_4 \neq 0$ and compensates for the lack of top-down control of attention from the high-level visual structures in the memorizing mode.

The coefficients a_1, a_2, a_3 and a_4 determine weights of the above items. They have been tuned experimentally using the following criteria: (i) the resulting scanpaths should cover the memorized images without cycling and (ii) the scanpaths should go through the semantically important image partitions if $a_4 \neq 0$.

Each shift of the AW ('eye movement') from nth fixation point is invariantly represented in the 'where' structure (motor memory) by the parameters of the mth context point which is selected to be the next $(n+1)$th fixation point. These parameters are defined with respect to the relative coordinate system X^1OY^1 in the current nth retinal image. The parameters of the shift are:

$$\psi_{\text{shift}}(n, n+1) = \psi_m(n);$$

$$l_{\text{shift}}(n, n+1) = \lambda_m(n);$$

$$\theta_{\text{shift}}(n, n+1) = l_{\text{o}}(n+1) - l_{\text{o}}(n), \qquad (16)$$

where $\psi_{\text{shift}}$and l_{shift} define the relative direction and relative distance of the shift, respectively; θ_{shift} defines the change of resolution in the central area of the AW when the latter moves from the nth to $(n+1)$th point of fixation.

Fig. 3(d) shows an example of shifting the AW to the next fixation point (the shift is shown by the arrow). The next fixation point is indicated in Fig. 4 by the cycle. The parameters of the shift are presented at the bottom of Fig. 4. During the memorizing mode, the model sequentially selects points of fixation and processes the image by shifting the AW along a scanpath of viewing (see Fig. 3(e) and (f)). During this process, the retinal images and AW shifts, related to the same image, are memorized in the sensory and motor memories, respectively and are alternatively connected forming the behavioral recognition program for this image.

In the recognition mode, the model checks the hypothesis about the object generated at the end of the search mode. In contrast to the memorizing mode, shifting the AW in the recognition mode is performed under the control of both top-down and bottom-up information. The bottom-up information depends on image content and contains the absolute parameters of the basic edge at the current nth point of fixation: orientation, $\varphi_{\text{o}}(n)$, and resolution, $l_{\text{o}}(n)$. The top-down information is recalled from the motor memory according to the executed behavioral program and the current AW position on the scanpath. This information contains the invariant (relative) parameters of the AW shift ('eye movement'): relative direction, $\psi_{\text{shift}}(n, n+1)$; relative distance, $l_{\text{shift}}(n, n+1)$, and relative resolution in the center of the AW, $\theta_{\text{shift}}(n, n+1)$. The AW controller uses both the above types of information for shifting the AW to the next location (fixation point).

3. Results

Gray-level images of scene objects and faces with the size 128 × 128 pixels and 256 gray levels were used in our test experiments. Examples of these images are shown in Fig. 5(a) and 6(a). In the memorizing mode, all tested images were sequentially presented to the model for memorizing. For 'semantically simple' images (scene objects, e.g. in Fig. 5(a)) the coefficient a_4 in Eq. (15) was set to zero, and the images were memorized without the 'semantic pre-tuning'. For the face images (e.g. in Fig. 6(a)), this coefficient was set to one. The value of the function of 'semantic significance' χ_{ij} was set to one for the image regions containing such semantically important image elements as 'eyes',

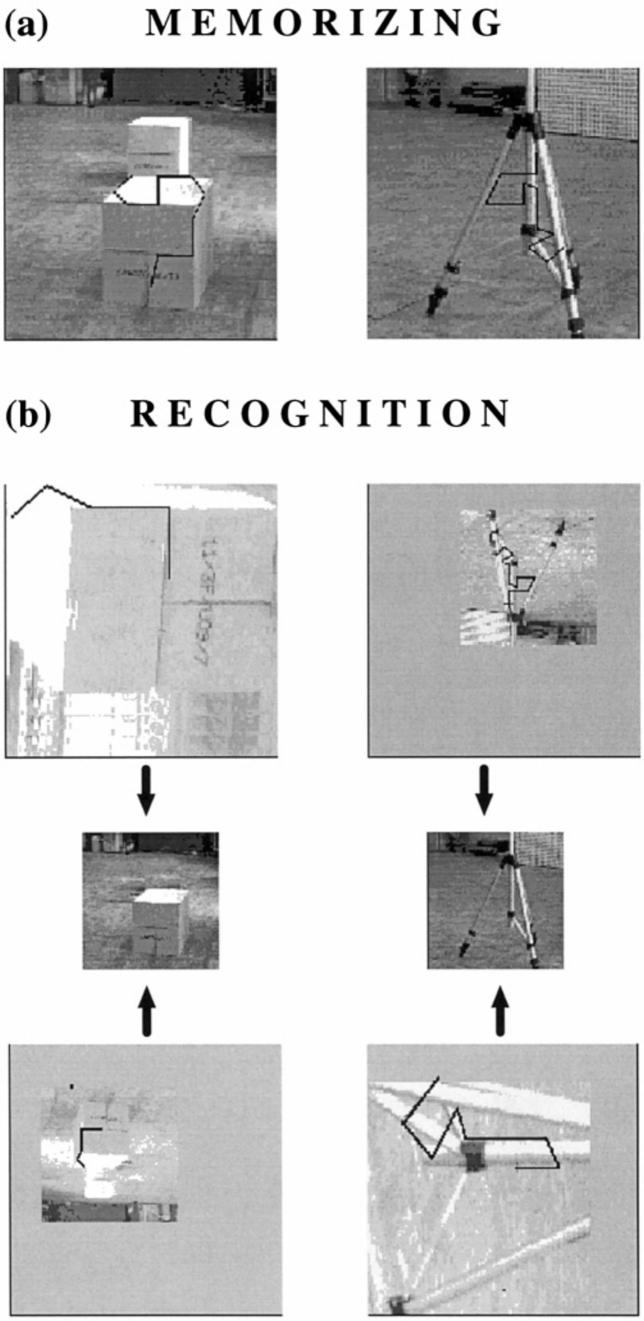

Fig. 5. Examples of recognition of test images invariantly with respect to shift, rotation and scale ('semantically simple' images). (a) The test images with the scanpaths in the memorizing mode. (b) The images presented for recognition (with the scanpaths during the recognition mode). These images were obtained from the test images by shifting, rotation and scaling. The results of recognition are shown by arrays.

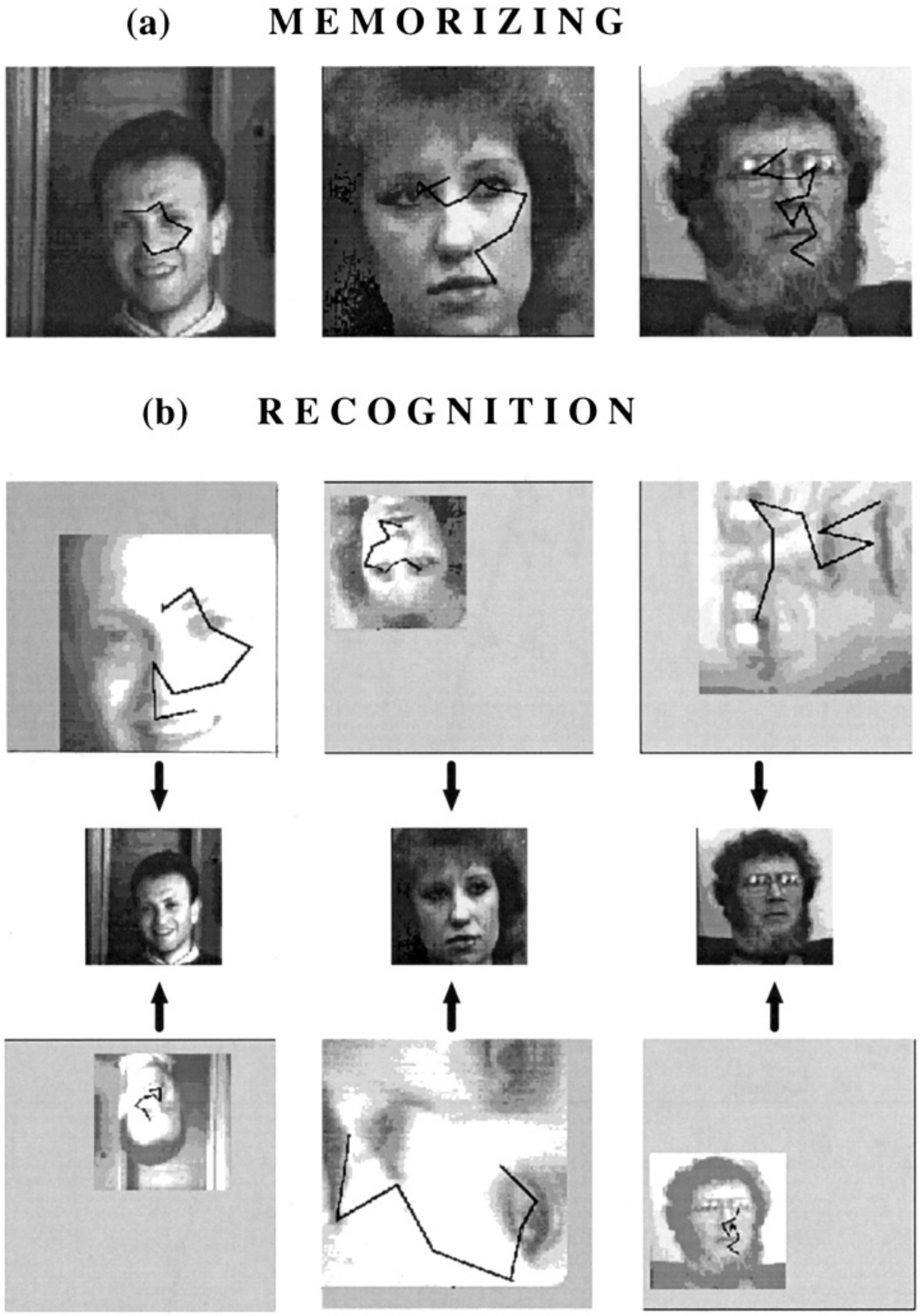

Fig. 6. Examples of recognition of test images (faces) invariantly with respect to shift, rotation and scale. For explanations see the legend to Fig. 5.

'mouth' and 'nose' and set to zero for the rest of the image.

In the memorizing mode, the model successively 'viewed' the images by way of sequential selection of fixation points and processing of the parts of the image within the AW centered at these points. The size of the AW varied depending on the selected resolution in the center of the AW. The representation of one 'retinal image' within the AW in the form of a set of the basic and context edges required 50 bytes of memory. The number of fixations per memorized image was 15–20 depending on the image complexity. Therefore, the complete representation of one image required about 1 K memory, which was 16 times less than the initial image representation. Thus, the model could memorize and store in memory a reasonably large number of images. We used an image database containing 20 images of simple scene objects and faces.

To test the model, we randomly selected one of the memorized images and used a special program which allowed us to shift the image within the raster of 128 × 128 pixels, resize it by 50–200%, and rotate it through any angle. A shifted, resized and rotated image was then presented to the model for recognition (see examples in Fig. 5(b) and 6(b)). Under these ideal conditions (i.e. without significant perturbations or big occlusions) the model usually demonstrated a 100% recognition. Occlusions or perturbations applied to a part of the image could break the recognition process at the scanpath points belonging to that part. However, the model was able to recognize an occluded/perturbed image by using the rest of the images if the latter contained a sufficient number of the scanpath points. Only significant occlusions or perturbations which broke the scanpath into short partitions (not allowing the required number of sequential successful matches) could cause an error when the model did not recognize the previously memorized image.

Thus, our test experiments showed that the model is able to recognize complex gray-level images invariantly with respect to shift, 2D rotation, and scale. The examples of invariant recognition of scene objects and faces are shown in Figs. 5 and 6, respectively. The scanpaths of image viewing in the memorizing mode are shown in Fig. 5(a) and 6(a). In the recognition mode, the model executes the corresponding behavioral program associated with the accepted hypothesis about the object. The scanpaths of viewing during the recognition mode are topologically similar to the scanpaths of viewing during the memorization of the same images (Fig. 5(b) and 6(b)).

4. Discussion

4.1. Model architecture: comparison with other models

The general architecture of our model is not new. A number of previous computational models of vision proposed a system architecture consisting of low- and high-level subsystems, with various top-down and bottom-up attention mechanisms controlling image memorizing and recognition [5,14,20–22,31–34,42,43]. These models underlined the importance of parallel-sequential (vs pure parallel) mechanisms for image processing, perception and recognition. They also indicated the significance of motor components in image representation, including explicit representation or emergent formation of scanpaths during image perception [5,14,20,22,42,43]. Some models utilized separated 'what' and 'where' high-level subsystems and considered this separation to be a fundamental property of the visual system [21,22,31–34,42]. These models focused on the deep analysis of different fundamental structural and functional aspects of visual recognition. However, in most cases the authors of these models did not present complete models which would be possible to test with respect to their ability to perform invariant recognition of real gray-level images. Some of these models were not tested in this context at all [14,42]. Other models dealt with relatively simple 2D images such as letters or binary objects without background [31–34,43]. It is not clear, whether these models are able to recognize objects in real gray-level images. Some robot and computer vision models were designed for the processing and recognition of objects in real scenes, but did not demonstrate an obvious ability to perform invariant recognition with respect to size and rotation [5,20–22]. Alternatively, the ultimate goal of our modeling efforts was to develop a biologically plausible model of the visual system capable of invariant recognition of real gray-level images.

4.2. Invariant image representation and recognition

The exciting ability of natural visual systems to perform invariant recognition has attracted the attention of scientists for more than a 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, invariant recognition is not an ultimate goal but rather a tool which helps the system to plan and execute actions adequate to the environment and consistent with the task at hand. This 'problem-oriented' (task-driven) behavior requires a 'problem-oriented' perception and recognition. According to Didday and Arbib [14], the goal is 'not to build a little internal copy of the visual scene, but rather to assist in making the decision of what action (if any) the organism should next initiate'. From this point of view, the 'absolute invariant recognition', when the visual system uses some special invariant transformation of the entire object's image and provides the same internal object representation (and hence the same output) at any object's location, size and orientation, is practically useless. In that case, the organism would 'learn' that the object is present in the scene, but would not 'know' how to manipulate it. For example, the system would know that a cup is present, but would 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 non-invariant object representation in memory (by transforming it to match the external image view) or manipulate 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, etc.

On the other hand, the complete lack of invariant representations makes the task of recognition practically unresolved. How would the system know which one of the majority of non-invariant 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 the two above opposite approaches.

A possible (and hopefully plausible) way for such combination may be based on invariant representation of an object's elements in the vicinity of fixation points (within the AW) and on the use of object manipulations (eye movements, shifts of the AW) in order to represent spatial relationships. We have tried to use this idea in the present model. The model holds invariantly represented image fragments in the sensory memory. Each fragment is associated with a certain object and with a certain action which the system should execute according to the behavioral recognition program corresponding to the object. In our model, the identification of a 'known' fragment (invariantly represented in memory) provides only a start to the recognition process which is executed under top-down control from the high-level modules. The recognition process, in turn, 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' algorithm 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.

The algorithm used in our model for 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. The idea of this algorithm is very simple and natural. 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 more coarse, generalized representation of a lager part of the image (or the entire image).

The algorithm used for invariant representation has no close analogues in previous computational models of vision. An alternative, biologically plausible algorithm, first described by Schwartz [43,44], was based on the logarithmic polar transformation. That algorithm also represents the image in a form invariant to size and rotation but has difficulties with the precise location of the fixation point. We have not found any experimental data that implicitly support or contradict our algorithm. Special psychophysical studies are needed to find out whether the visual system can use relative orientations and/or relative angular locations of edges for the invariant image representation and recognition.

The question of whether or not the natural visual system uses invariant image transformations is not completely understood. For example, it was shown that the recognition time in humans increased with the angle of rotation [45] and size scaling [46] of the recognized images. These data support the idea that the visual system does not use invariant transformations and representations, but recognizes rotated and resized images using mental rotation and scaling operations. On the other hand, a dichotomy between object recognition and mental imagery and the ability of human subjects to perform a fast rotation-invariant recognition have been also demonstrated [47,48]. Thus, the question of whether or not the mental rotation and scaling operations are essential for invariant recognition (whether they provide key or additional mechanisms) is still open. It is also important to note that the time of recognition in psychophysical experiments is usually measured from the moment of image presentation to the moment of making a decision about the object [45,46]. This overall recognition time actually depends on many factors (type and complexity of the image, existence of context used as references, pretuning of human subjects, etc.) and is a complex composition of the durations of a number of functionally different subprocesses (search/viewing, focusing/foveation, generating a right/wrong hypotheses about the object, testing the hypotheses, etc.). Thus, a further careful consideration of all relevant factors and evaluation of the durations of different subprocesses are probably required in order to use temporal characteristics for a principal validation of mechanisms for invariant recognition.

Our model is based on the idea that image fragments are represented in memory in an invariant form and the time of their recognition therefore does not depend on their orientations. The overall object recognition time in our model contains two phases: the phase of search which ends with generating a hypothesis about the object, and the phase of actual recognition when the system examines the hypothesis using top-down driven shifts of attention.

The duration of the search phase depends on the scanning algorithm used, location and size of the image, the number of memorized images, and the number of fixations used in the memorizing mode (the more fixations, the faster search). We expect, that the duration of the search phase should be significantly reduced by the problem-oriented, top-down mechanisms which were not considered in the current model.

The duration of the recognition phase (recognition mode) in our model depends neither on the object location, size and orientation nor on the number of memorized images. It is defined by the number of sequential successful matches which is considered to be enough for making a decision on recognition. This number is limited by the overall number of fixations used during memorization of the image. We set this number in advance (dependent or independent on overall numbers of used fixations), usually about six to seven. In fact, this number defined a 'psychological self-confidence' of the system. Sometimes, the accepted hypothesis was not confirmed at some fixation point in the recognition mode (the predicted image fragment did not correspond to the current one) and the system returned to the previous phases which prolonged the overall recognition time. This was dependent on how similar the memorized images were and how many fixation points were selected within the similar parts of different images during their memorizing.

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 is able to recognize these objects in multi-object scenes with complex backgrounds. In contrast, the natural visual system uses special mechanisms which 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.; [6,13,28,49]). Additional mechanisms, which separate the objects in the image from each other and from the background, should be incorporated into the model to allow memorizing objects in complex multi-object images. These mechanisms will prevent a selection of fixation points outside the object of interest.

In conclusion, our model provides important insight into the role of behavioral aspects for invariant pattern recognition. Some model predictions of human vision mechanisms await special psychophysical experimental investigations. These include, for example: an invariant encoding of image elements projected onto the fovea; the use of elementary features (e.g. edges) in the fovea center as a basis (frame of reference) for invariant encoding of both the retinal image and the succeeding eye movement; the dependence of the selected location of the next fixation point on the parameters of the basic features in the fovea. The basic algorithmic ideas of the model and approach used may be applied to computer and robot vision systems aimed toward the invariant image recognition (for example, see Ref. [50]).

References

[1] Cowey A. Projection of the retina onto striate and parasrtiate cortex in the squirrel monkey Saimiri Sciureus. J Neurophysiol 1964;27:266–393.

[2] Stone J, Fukuda Y. Properties of cat retinal ganglion cells: a comparison of W-cell with X- and Y-cells. J Neurophysiol 1974;37:722–48.

[3] Wilson JR, Sherman SM. Receptive-field characteristics of neurons in cat striate cortex: changes with visual field eccentricity. J Neurophysiol 1976;39:512–33.

[4] Van Essen D. Functional organization of primate visual cortex. In: Peters A, Jones EG, editors. Cerebral Cortex. New York: Plenum, 1985:259–329.

[5] Burt P J. Smart sensing within a pyramid vision machine. Proc IEEE 1988;76:1006–15.

[6] Julesz B. Experiments in the visual perception of texture. Sci Am 1975;232:34–43.

[7] Julesz B. A brief outline of the texton theory of human vision. Trends Neurosci 1984;7:41–5.

[8] Neisser V. Cognitive Psychology. New York: Appleton, 1967.

[9] Noton D, Stark L. Scanpaths in eye movements during pattern recognition. Science 1971;171:72–5.

[10] Posner MI, Presti DE. Selective attention and cognitive control. Trends Neurosci 1987;10:13–7.

[11] Shiffrin RM, Schneider V. Controlled and automatic human information processing. 2. Perceptual learning, automatic attending and a general theory. Psychol Rev 1977;84:1270–90.

[12] Treisman AM, Gelade G. A feature integration theory of attention. Cogn Psychol 1980;12:97–136.

[13] Yarbus AL. Eye Movements and Vision. New York: Plenum, 1967.

[14] Didday RL, Arbib MA. Eye movements and visual perception: a two visual system model. Int J Man-Machine Stud 1975;7:547–69.

[15] Kosslyn SM, Flynn RA, Amsterdam JB, Wang G. Components of high-level vision: a cognitive neuroscience analysis and account of neurological syndromes. Cognition 1990;34:203–77.

[16] Ungerleider LG, Mishkin M. Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW, editors. Analysis of Visual Behavior. Cambridge, MA: MIT Press, 1982:549–86.

[17] Mishkin M, Ungerleider LG, Macko KA. Object vision and spatial vision: two cortical pathways. Trends NeuroSci 1983;6:414–7.

[18] Bajcsy R. Active perception. Proc IEEE 1988;76:996–1005.

[19] Ballard DN. Reference frames for animate vision. In: Proceedings of 11th International Joint Conference on Artificial Intelligence. Detroit: Morgan Kauffmann, 1989:1635–1641.

[20] Bolle RM, Califano A, Kjeildsen R. Data and model driven foveation. IBM Technical Report, 1989.

[21] Rao RPN, Ballard DN. Dynamic model of visual recognition predicts neural response properties in the visual cortex. Technical Report 96.2. University of Rochester, 1996.

[22] Rao RPN, Zelinsky GJ, Hayhoe MM, Ballard DN. Eye movements in visual cognition: a computational study. Technical Report 97.1. University of Rochester, 1997.

[23] Rimey RD, Brown CM. Selective attention as sequential behavior: Modeling eye movements with an augmented hidden Markov model. Technical Report 327. University of Rochester, 1990.

[24] Hubel DH, Wiesel TN. Receptive fields, binocular integration and functional architecture in the cat's visual cortex. J Physiol 1962;160:106–54.

[25] Hubel DH, Wiesel TN. Sequence, regularity and geometry of orientation columns in the monkey striate cortex. J Comp Neurol 1974;158:267–93.

[26] Enns JT, Rensink RA. Influence of scene-based properties on visual search. Science 1990;247:721–3.

[27] Enns JT, Rensink RA. Preattentive recovery of three-dimensional orientation from line drawing. Psychol Rev 1991;98:335–51.

[28] Marr D. Vision. New York: WH Freeman, 1982.

[29] Palmer SE. The psychology of perceptual organization: a transformational approach. In: Beck J, Hope B, Rosenfeld A, editors. Human and Machine Vision. New York: Academic Press, 1983.

[30] Hinton GE, Lang KJ. Shape recognition and illusory conjunctions. Procedings of 9th International Joint Conference Artificial Intelligence. Los Angeles, 1985:252–259

[31] Carpenter GA, Grossberg S, Lesher GW. A what-and-where neural network for invariant image preprocessing. Proceedings of International Joint Conference on Neural Networks, Piscataway, NJ: IEEE Service Center. 1992:303–308.

[32] Olshausen B, Anderson C, Van Essen D. A neural model of visual attention and invariant pattern recognition. CNS Memo 18, Pasadena: California Institute of Technology, 1992.

[33] Otto I, Grandguillaume P, Boutkhil L, Burnod Y. Direct and indirect cooperation between temporal and parietal networks for invariant visual recognition. J Cog Neurosci 1992;4:35–57.

[34] Rueckl JG, Cave KR, Kosslyn SM. Why are 'what' and 'where' processed by separate cortical visual systems? A computational investigation. J Cog Neurosci 1989;1:171–86.

[35] Grossberg S, Mingolla E, Todorovic D. A neural network architecture for preattentive vision. IEEE Trans Biomed Eng 1989;36:65–84.

[36] Lippmann RP. An introduction to computing with neural nets. IEEE Trans ASSP, 1989;4:4–22.

[37] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 1982;79:2554–8.

[38] Cave KR, Wolfe JM. Modeling the role of parallel processing in visual search. Cogn Psychol 1990;22:225–71.

[39] Wolfe JM, Cave KR. Deploying visual attention: the guided search model. In: Troscianko T, Blake A, editors. AI and the Eye. Chichester, UK: Wiley, 1990:79–103.

[40] Wolfe JM. Guided Search 2.0: a revised model of visual search. Psychon Bull Rev 1994;1:202–38.

[41] Posner MI, Cohen Y. Components of visual orienting. In: Bouma H, Bowhuis D, editors. Attention and Performance X. Hillsdale, NJ: Erlboum, 1984:531–56.

[42] Rimey RD, Brown CM. Selective attention as sequential behavior: Modeling eye movements with an augmented hidden Markov model. Technical Report 327. University of Rochester, 1990.

[43] Yeshurun Y, Schwartz EL. Shape description with a space-variant sensor: algorithms for scan-path, fusion, and convergence over multiple scans. IEEE Trans Pattern Anal Mach Intell 1989;11:1217–22.

[44] Schwartz EL. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. Brain Res 1980;20:645–69.

[45] Tarr MJ, Pinker S. Mental rotation and orientation-dependence in shape recognition. Cogn Psychol 1989;21:233–82.

[46] Larsen A, Bundesen C. Size scaling in visual pattern recognition. J Exp Psychol: Hum Percept Perform 1978;4:1–20.

[47] Farah MJ, Hammond KM. Mental rotation and orientation-invariant object recognition: dissociable process. Cognition 1988;29:29–46.

[48] McMullen PA, Farah MJ. Viewer-centered and object-centered representations in the recognition of naturalistic line drawings. Psychol Sci 1991;4:275–7.

[49] Gibson JJ. The Perception of the Visual World. Boston: Houghton Miflin, 1950.

[50] Hecht-Nielsen R, Zhou YT. VARTAC: A foveal active vision ATR system. Neural Netw 1995;8:1309–21.