Saccadic Object Recognition with an Active Vision System [*]

G.–J. Giefing H. Janßen H. Mallot

Institut für Neuroinformatik, Ruhr–Universität Bochum
44780 Bochum, Germany
e–mail heja@neuroinformatik.ruhr-uni-bochum.de

Abstract

We propose an active vision system for saccadic camera gaze shifts and explorative scene analysis as a new integral approach to image understanding. The model consists of two sensory subsystems: preattentive peripheral feature detection and high resolution foveal image identification based on a hypercolumnar representation. Visual objects are non–explicitly stored in two sparsely coded associative memories separating fixation locations from identities of foveal views. An egocentric interest map integrates bottom–up and top–down information sources and decides when to generate a camera movement. A selective masking of preattentive processes supports a cooperation with cognitive object recognition. The system is easily extendible, copes with occlusions and distortions and can be driven in different modes for exploration tasks. This model is able to perform visual search and reproduce findings in the human visual system.

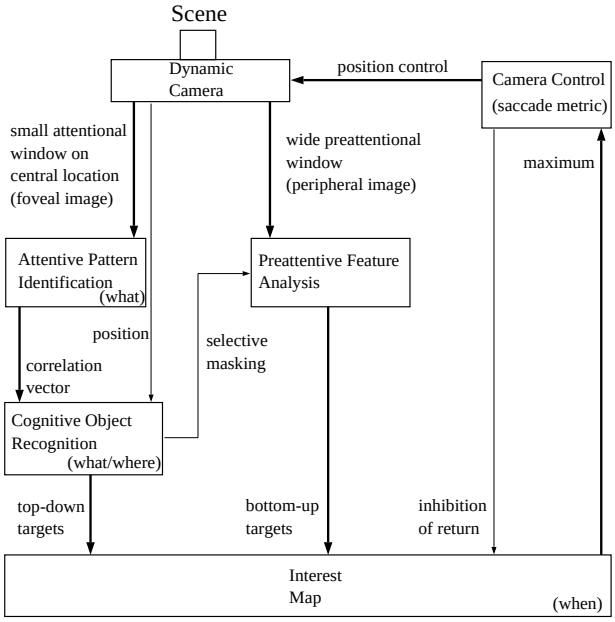

Figure 1: System survey.

1 System Survey

In this paper we propose an *animate image understanding system* for static two–dimensional scene analysis and behavioural object recognition by means of explicit camera gaze shifts. The function of the system is designed by a hierarchy of physiological and anatomical motivated structures (i.e., fovea, hypercolumns, associative memory).

The scene is supposed to be a mainly static external memory buffer. An *implicit selective attention mechanism* which is considered as a behavioural context or system status modulates perception by controlling information flow and distributing processing resources to scene locations.

A space–variant area magnification factor of a mapping based *scale–sized fovea model* [4] provides a small central area of high resolution coinciding the camera gaze direction and a large low resolution periphery.

Loci of peripheral spatiotemporal features discontinuities are detected by a *preattentive feature analysis*.

By foveating these spatial positions (*bottom–up saccades*) an *attentive identification process* analyzes fine characteristic details. For the pattern identification we use a pattern recognition subsystem evaluating local orientations which supplies a correlation histogram based on a large set of stored foveal patterns [3]. It is invariant to restricted translations and distortions. For reasons of simplicity the location of the attentive processing window is always chosen identically to the foveal area (*overt attention*). Thus the system is forced to shift the camera gaze for an object recognition task.

The cognitive representation of objects is constructed of several foveal views whose attentive iden-

[*]Edited version, originally published in: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 664 – 667, IEEE (1992)

[†]Supported by the German Federal Department of Research and Technology (BMFT), Grant No. ITR8800K4

tification information, preattentive detection information ("what") and spatial position ("where") are stored in two separate memory sections (*"what–where"–separation* [8]). An interactive supervised learning procedure guided by the peripheral preattentive feature detection module selects conspicuous locations of an object (*supervised preattentive learning*). The foveal patterns of these locations including the excitation vector of the preattentive feature detectors and their object centered spatial coordinates are put into the memories.

A *cognitive object recognition process* integrates transsaccadic information to state hypotheses about the scene and tries to verify object hypotheses by claiming saccades to scene locations specific for an object (*top–down saccades*). A *selective masking* of the preattentive pathway enhances the performance of the search for the peripheral feature or feature conjunctions of the desired scene location. This part of the selective attention mechanism enforces the cooperation between the two pathways improving estimations of spatial gaze positions.

The attentive and preattentive processing pathways feed their saccadic target demands [9] into the *excitatory* part of the egocentric *interest map*. The interest map covers the whole scene on the resolution scale of the camera position precision and uses a spherical coordinate system centered in the actual direction of gaze. This map implements competition and cooperation of the different demands and also decides "*when*" to saccade (synchronisation). The *inhibitory part* avoids a return to positions already foveated. We propose the interest map as a *central information integration facility* able to incorporate future modules (e.g. smooth pursuit camera movements, depth information). The description of the map is continuously in time. All target positions are "forgotten" by intrinsic diffusion and relaxation processes. The processing structure of the system is depicted in Fig. 1.

2 Preattentive Feature Analysis

A foveal compression by a foveal mapping $\mathcal{R} : \mathbb{R}^2 \mapsto \mathbb{R}^2,\ \mathbf{y} \mapsto \mathbf{x}$ transforms the image I from view centered image coordinates $\mathbf{y}$ to retinal sampling coordinates $\mathbf{x}$ according to a desired ganglion cell distribution.

Because of the mainly static scene we evaluate the discontinuities of simple peripheral features in space only. The detection of temporal changes in our preattentive model is not feature specific. In fact, saccades of humans to structured targets are easier to predict if

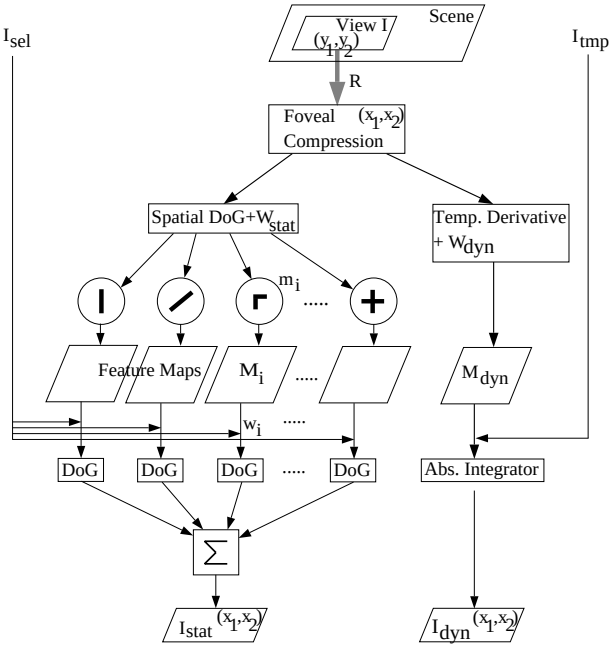

Figure 2: Scheme of our preattentive module (only one spatial frequency channel c_l is shown).

it is a (transient) on–set and not a sustained stimulus (H. Deubel, pers. comm.).

2.1 Sustained Preattentive Pathway

In the sustained pathway we feed $I(\mathbf{x})$ at the end t_k of fixation k into different spatial frequency channels c_l to become invariant of illumination and to analyze the signal on different scales σ_l. Peripheral information is space–variantly enhanced by W_{stat} to force gaze shifts.

$$c_l(\mathbf{x}, t_k) \;=\; W_{stat}\, \nabla^2 G_{\sigma_l}(\mathbf{x}) \;*\; I(\; \mathcal{R}(\mathbf{y}), t_k \;)$$

A feature map M_{il} represents the cross–correlation of channel output c_l and feature m_i:

$$M_{il}(\mathbf{x}, t_k) \;=\; m_i(\mathbf{x}) \;\otimes\; c_l(\mathbf{x}, t_k).$$

We detect as features m_i simple lines, curved contours and crossings similar to those found to be processed in cortical area V1. Steeply descending autocorrelation functions of the features improve the precision of peripheral feature localization.

A Marr–Hildreth operator computes discontinuities (e.g., texture boundaries, occluding objects) in the feature maps M_{il} on a coarse scale σ. Prior to a final summation to get I_{stat}, the cognitive recognition process masks these channels by $I_{sel} = (w_{il})$ supporting the search for a specific feature or feature conjunction (C.M. Brown, pers. comm.).

$$I_{stat}(\mathbf{x}, t_k) \;=\; \sum_{il}\; w_{il}\;\; \nabla^2 G_\sigma \;*\; M_{il}(\mathbf{x}, t_k)$$

2.2 Transient Preattentive Pathway

A reflexive behaviour for temporal events (arousal) is the task of a featureless transient analysis during a fixation interval between two saccades $[t_{k-1}, t_k]$. This detector integrates space–variantly weighted (W_{dyn}) temporal changes. The cognitive pathway has the possibility to (space–invariantly) mask the transient detection by I_{tmp} to switch into a purely static scene analysis.

$$I_{dyn}(\mathbf{x}, t_k) \;=\; \int_{t_{k-1}}^{t_k} \left| \; I_{tmp}\; W_{dyn}(\mathbf{x})\; \frac{\partial}{\partial t}\; I(\mathbf{x}, t) \; \right| \; dt$$

At the end t_k of a segregated fixation period I_{stat} and I_{dyn} are added to our interest map. A scheme of the preattentive module is shown in figure 2.

2.3 Salient Feature Detection for Visual Search

This module is the base of an explorative saccadic scanpath without any high–level object knowledge. Because of the spatial integration property of the $\nabla^2 G$ operator the preattentive target computation shows a *global* or *peripheral effect* known from saccadic eye-movement recordings to structured targets.

The sustained preattentive pathway, as a matched filter, is able to select immedeately the location $\mathbf{x}^*$ of a feature or feature conjunction which holds the maximum $I_{stat}(\mathbf{x}^*) \;\geq\; I_{stat}(\mathbf{x})$. With the help of cognitive masking, it is always possible to find a certain feature location which differs from its surrounding in at least one feature dimension (*salient feature*). This excellently agrees with results of visual search [7]. Overlapping detector channels and direct summation of their outputs show the continuous properties of *target–nontarget* and *nontarget–nontarget similarities* [1].

3 Cognitive Object Recognition

3.1 Generation of an Object Hypothesis

The identification process supplies a *correlation vector* $\mathbf{c}$ between the current foveal image and all stored foveal patterns at the end t_k of fixation k. The maximum value[1]

$$c_{i'}(t) = \max_i(\mathbf{c}_i(t))$$

determines the actual recognized pattern i'.

The cognitive object recognition process integrates temporally transsaccadic information in a vector we call "object accumulator" $\mathbf{a}(t)$:

$$\frac{d\mathbf{a}(t)}{dt} \;=\; \underbrace{\mathbf{Q}\mathbf{c}(t)}_{\text{input}} \;-\; \underbrace{\mathbf{d}(t)\cdot\mathbf{a}(t)}_{\substack{\text{position}\\ \text{error}}} \;-\; \underbrace{\tau_a \mathbf{a}(t)}_{\text{relaxation}}$$

An internal object hypothesis for object j' is stated if

$$a_{j'} = \max_j(a_j) \;\wedge\; a_{j'} > \theta$$

is valid, where θ is a threshold to inhibit the statement of 'weak' hypotheses.

The *input term* is a vector containing the significance of the current foveal image to all the v learned objects.

The degree of membership of the known u foveal patterns to the v objects is stored in the pattern–object relation matrix $\mathbf{Q} \in \mathbb{R}^{u\times v}$. The N preattentive detector responses for a modulated visual search are held in $\mathbf{S} \in \mathbb{R}^{N\times u\times v}$. $\mathbf{Q}$ and $\mathbf{S}$ constitute the associative "what"–memory.

The "where"–part stores the position of the foveal patterns of an object in an association matrix $\mathbf{R} \in \mathbb{C}^{u\times v}$, which keeps the relative position (as a complex number) of all u pattern in all v objects. Only the value $a_{j'}$ of the actual object hypothesis is diminished by the *position error term* $\mathbf{a}\cdot\mathbf{d}$. The latter depends nonlinearly on the spatial difference between the last really executed saccade (position $r'(t_{k-1})$ to $r'(t_k)$) and the difference of the last two recognized patterns ($i'(t_{k-1})$ and $i'(t_k)$) according to the "where'–memory (position $r_{i'(t_{k-1})j'(t_k)}$ to $r_{i'(t_k)j'(t_k)}$):

$$d_j(t) = \begin{cases} f(\Delta r) = \Delta r^2 & \text{for} \quad j = j' \\ 0 & \text{for} \quad j \neq j' \end{cases} \qquad \text{where}$$

$$\Delta r = (r'(t_k) - r'(t_{k-1})) - \left(r_{i'(t_k)j'(t_k)} - r_{i'(t_{k-1})j'(t_k)}\right).$$

To agree with psychophysical findings $O(f(\Delta r)) = o(\Delta r)$ must be valid.

The *relaxation term* enables the system to "forget" acquired information.

$\mathbf{Q}, \mathbf{R}, \mathbf{S}$ will be very sparse in case of realistic numbers of foveal patterns and objects. Because this *emergent saccadic model* uses a very efficient and robust representation avoiding a memory expensive high resolution scene buffer, it does not rely on any specific or explicit scanpath for recognizing an object [6]. Although it will generate an object specific scanpath for a known object, if the sensory information is similar to that during the learning steps.

3.2 Generation of Top–Down Targets

In a structure called *pattern accumulator* $\mathbf{b}$, the cognitive object recognition module processes the pattern identification information for object j'. The relative size of the values in $\mathbf{b}$ denote the urgency to foveate a

[1] For reasons of simplicity, we choose a continuous notation.

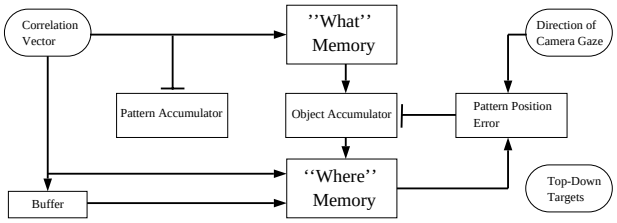

Figure 3: Object hypothesis generation

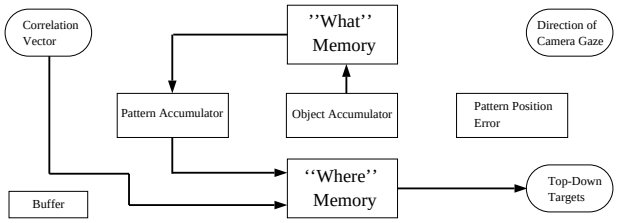

Figure 4: Cognitive target generation

pattern to verify the current hypothesis while trying to avoid patterns, which have lately been gazed at:

$$\frac{d\mathbf{b}(t)}{dt} = - \underbrace{\mathbf{c}(t)}_{\text{recognition}} + \underbrace{\mathbf{Q}^T \mathbf{a}'(t)}_{\text{verification}} - \underbrace{\tau_b \mathbf{b}(t)}_{\text{relaxation}}$$

$$\text{where} \qquad a'_j(t) = \begin{cases} a_j(t) & \text{for } j = j' \\ 0 & \text{for } j \neq j'. \end{cases}$$

A *recognition term* diminishes all values b_i by the significance they have already been recognized with.

The *verification term* is calculated by the backprojection of the current object hypothesis j' according to the matrix $\mathbf{Q}$. That is why without an object hypothesis all values b_i for patterns already seen become largely negative.

Values of $\mathbf{b}$ are slowly "forgotten" by a temporal *relaxation*.

By using the "where"–memory, the system generates a weighted list of discrete top–down target positions $\mathbf{e}$ of object j' which is transformed into a smooth excitation distribution I_{obj} for the continuous interest map.

$$\mathbf{e}(t) = (\mathbf{R}^T \, \mathbf{a}'(t), \, \mathbf{b}(t))$$

The most urgent pattern p' which holds

$$b_{p'}(t) = \max_p(\mathbf{b}(t))$$

masks the preattentive processing module using $\mathbf{S}$ by

$$I_{sel} = (s_{j'p'})$$

If $a_{i'}$ exceeds a threshold η, so that $a_{i'} > \eta > \theta$, temporal changes in the view are completely ignored by the system assigning $I_{tmp} = 0$. The data flow for this process is depicted in figure 4.

4 Interest Map and Camera Control

The bottom–up and top–down subsystems register their information after an appropriate coordinate transform in the *excitatory section* I_{exc} of the *interest map*:

$$\begin{aligned} \frac{dI_{exc}(t)}{dt} &= I_{stat} + I_{dyn} + I_{obj} + D_{exc}\Delta I_{exc} \\ &\quad -(1-\tau_{exc})I_{exc}(t)\Big|_{view} - \tau_{exc} I_{exc}(t) \\ \frac{dI_{inh}(t)}{dt} &= I_{cam} + D_{inh}\Delta I_{inh} - \tau_{inh} I_{inh}(t) \end{aligned}$$

The *camera control* enters I_{cam} into the *inhibitory section* I_{inh} to prevent the gaze from a return to already fixated positions. Thus the system shows an *instable behaviour* with respect to gaze positions.

Since one view does not cover the scene completely, the system is not able to notice all environmental changes. Therefore we introduce a relaxational term τI in the two sections of the map which allows to again foveate locations after some time. A spatial diffusion $D\Delta I$ locally distributes the activity over the map, so the camera control system can cope with the integration of positional errors.

The camera control can easily calculate the spatial position $\mathbf{z}^*$ of the next target holding for the maximum of the sum of the two sections $I_{exc}(\mathbf{z}^*) + I_{inh}(\mathbf{z}^*) \geq I_{exc}(\mathbf{z}) + I_{inh}(\mathbf{z})$ after all entries are done.

This map is used to synchronize processes running on different workstations in parallel and may decide *when* to evocate event–triggered saccades also. For more advanced exploration, we are currently evaluating control strategies prefering an average saccadic distance and an inhibition of subsequent saccades following the same direction.

5 Implementation and Results

The system is able to recognize two–dimensional objects [2]. Tests show that this object recognition

system can cope with different kinds of difficulties: imperfect patterns, slightly distorted positions of saccadic fixation points, occlusions and added object parts (distractions). A switching of hypothesis is also possible by presenting a scene with two objects sharing several local views. A quantitative performance evaluation will be published soon.

Additionally we are working on a closer interaction of preattentive and attentive processing and the integration of stereo based depth information for 3–D recognition.

References

[1] J. Duncan and G. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433 – 458, 1989.

[2] G.-J. Giefing, H. Janßen, and H. A. Mallot. A saccadic camera movement system for object recognition. In *Proc. ICANN–91*. Elsevier North–Holland, 1991.

[3] H. Janßen and J. Kopecz. Image representation and associative recognition in hypercolumnar scale space structure. this volume, 1991.

[4] H. A. Mallot and G.-J. Giefing. Retinal sampling grids and space-variant image processing. In *Parallel Processing in Neural Systems and Computers*, pages 125 – 128. DGK, North-Holland, 1990.

[5] G. Palm. On associative memory. *Biol. Cybern. 36, 19-31*, 1987.

[6] R. D. Rimey and C. M. Brown. Selective attention as sequential behavior: Modeling eyemovements with an augmented hidden markov model. Technical Report TR327, Computer Science Department, University of Rochester, 1990.

[7] A. Treisman. Preattentive processing in vision. *Computer Graphics and Image Processing*, 31:156 – 177, 1985.

[8] L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *The Analysis of Visual Behavior*, pages 549 – 586. The MIT Press, Cambridge, Ma., 1982.

[9] P. Vivani. Eye movements in visual search: cognitive, perceptual and motor control aspects. In E. Kowler, editor, *Eye Movements and Their Role in Visual and Cognitive Processes*. Elsevier Science Publishers B.V., 1990.