S. Fuchs R. Hoffmann (Hrsg.)

Mustererkennung 1992

14. DAGM-Symposium
Dresden, 14.-16. September 1992

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

A Neural Network Hierarchy for Data Driven and Knowledge Controlled Selective Visual Attention*

H.-M. Gross, R. Franke, H.-J. Boehme, Claudia Beck

Technical University of Ilmenau
Department of Neuroinformatics
O-6300 Ilmenau, P.O.B. 327, Germany
email: gross@informatik.th-ilmenau.de

Abstract

We present a neural implementation of a dynamical network hierarchy for data driven and knowledge controlled selective visual attention. The model architecture is composed of several interacting subsystems for different processing tasks. With the example of real-world scene analysis the proposed model demonstrates its abilities in preattentive search and in decomposition of a complex visual input into a sequence of striking local input segments. Based on its functional architecture our model is able to shift its focus of attention both driven by the input data and controlled by its internal processing state and the already acquired knowledge.

1. Introduction and Model Hypothesis

The phenomenon of selective attention in human visual perception points the way out of the dilemma of combinatorical explosion in analysis of real-world visual scenes: there are sequential processing modes intermingled with the parallel one. Selective attention means breaking down the flow of information too high to be managed by the analyzing system in parallel into meaningful pieces of lower dimension. The benefit of selective visual attention is, that the analyzing or identifying system has not to deal with all visual inputs in parallel but only with a limited sequence of presorted, lower dimensional groups of input elements [1]. This way the analyzing system can focus attention on the most desired visual stimulus among several simultaneously active stimuli, both driven by the input data and controlled by its internal processing state (hypothesis about the input data) and the already acquired knowledge. This control of the attentive search during the recognition process is a fundamental mechanism for self-organization of sequential and episodical representations, a special type of non-trivial dynamical knowledge about spatio-temporal processes [2].

Our model concept and the implemented mechanisms have been influenced essentially by the neurophysiological concepts of primary visual processing and selective attention of Koch [3] and Orban [4]. Koch assumes that selective visual attention operates on a set of topographical cortical maps encoding the visual environment. This early representation includes a variety of maps for different elementary features such as orientation of edges, textural features, color, disparity etc. This multiparametrical preprocessing and mapping is confirmed by Orban [4]. He showed that in the first sensory visual area instead of a hierarchical description of the input features a very rich parallel representation of the input by parameter filters is done. In order to simulate the preattentive search and the interaction with the attentive mode, we had to implement special mechanisms in our model

*Supported by the German Federal Department of Research and Technology (BMFT), Grant No. 413-5839-01 IN 101D - NAMOS-Project

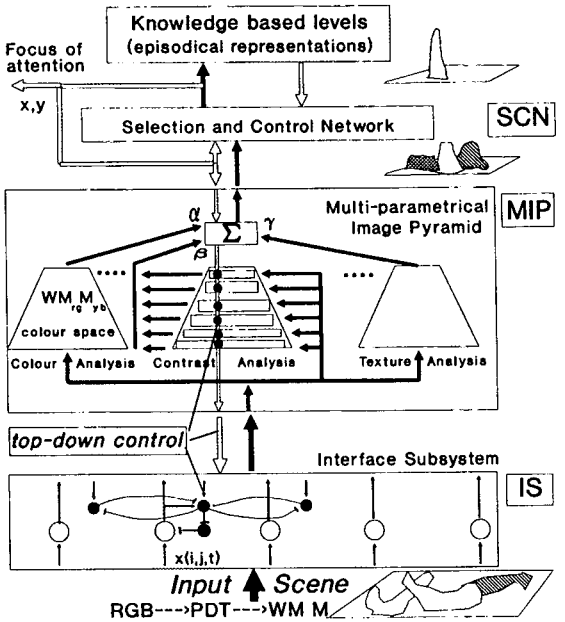

Figure 1: *Simplified functional architecture of the network hierarchy and the main subsystems*

- to yield a measure of the conspicuity of a location in the visual scene
- to select the most active area (many different feature detectors are simultaneously active) in a mapping plane where the different feature maps are superimposed
- to shift the focus of attention from the current to the next striking location in the scene
- to control and manipulate the preattentive flow of information in the course of an attentional process taking into account internal system's knowledge about the visual scene.

2. Functional Architecture of the Network Hierarchy

Our model architecture is composed of various interacting dynamical subsystems for different processing tasks. All these multi-layered neural subsystems constituting a dynamical control hierarchy are strongly interrelated by information and control streams. A strongly simplified scheme of the interrelations and the functional architecture of the main subsystems

- Interface Subsystem - **IS**
- Multi-parametrical Image Pyramid - **MIP**
- Selection and Control Network - **SCN**

is shown in Figure 1. It is to note, that despite the different model hypothesis and structural implementations of our approach compared to the active vision system in [5] some functional aspects of preattentive and attentive control of saccadic image scanning are similar. Detailed aspects of the real-world scene pre-processing and the data transformation between the several colour spaces are not discussed in this paper. Instead of this only a short overview about the implemented transformations

between the colour spaces will be given in the following. Starting point is the RGB-image of the real-world scene, that is transformed in a first step into a PDT image, a special type of a three-dimensional colour space. This transformation is based on the Hering red/green and yellow/blue opponent systems, a special concept of trichromacy, which is clearly validated by neurophysiological results [6]. The PDT colour space is comparable to the XYZ standard colour space, that is also based on the opponent colour theory and is in agreement with the traditional view of colour scientists. Because of its neurophysiological plausibility this PDT data set then is scaled logarithmically. Finally, based on Luther's transformation, the PDT image is transferred into the $WM_{rg}M_{yb}$ space, a colour space that is better suited for the following analysis in the different feature extraction pathways of our model. The three components of this space are the black-white process (intensity or brightness) $W = P$, the red/green opponency $M_{rg} = D - P$ and the yellow/blue opponency $M_{yb} = P - T$. All intensity (activity) based mechanisms of our model use only the W component of this colour space, while the colour analyzing pathway in the Multi-parametrical Image Pyramid (see 2.2) uses all dimensions of the $WM_{rg}M_{yb}$ space.

2.1 The Interface Subsystem

To our mind selective visual attention requires the freedom to select only those parts out of the input which are needed at that time in the analyzing or recognition process. Therefore we have implemented in our hierarchy the controllable **Interface Subsystem - IS**, which is based on a simplified model architecture of the thalamo - reticular complex proposed in [1]. Operating on this Interface the succeeding higher subsystems can actively manipulate the bottom-up flow of input data streams giving them the freedom of formulating and testing hypothesis on interesting parts of the parallel input. This way they are able to control the interface and to perform a random access grouping at the interface level to restrict the complexity of the input to that needed at this time.

2.2 The Multi-parametrical Image Pyramid

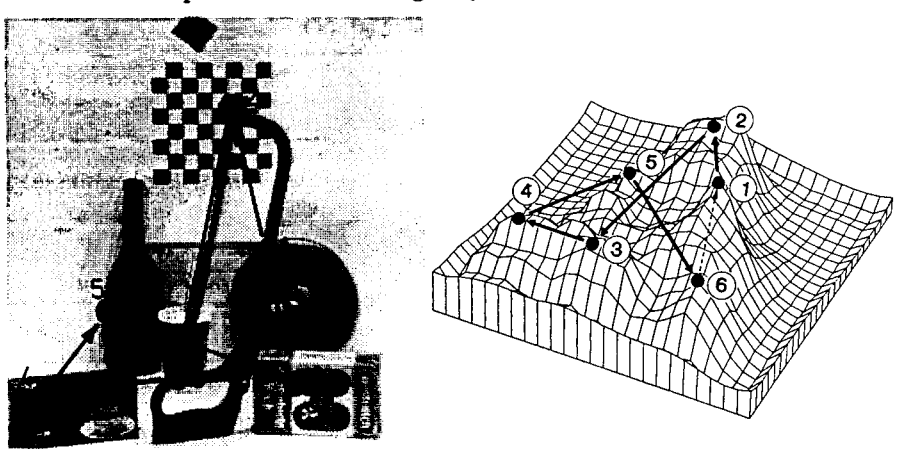

Figure 2: *(Left side) Input scene for simulation the data driven preattentive search. The marked scan-path (1-6) shows the course of the preattentive search and the sequence of selected most striking input locations. (Right side) Sequential selection of the most active peaks within the activity landscape of SCN and shifting the focus of attention between the most striking locations in the scene (see 3.).*

The **Multi-parametrical Image Pyramid - MIP** operating on the Interface Subsystem detects in distinct analyzing pathways differences in local conspicuous features of the IS activity pattern

(Fig. 2 - left side). By reducing the resolution and size of the several processing levels we get a processing pyramid that realizes both an enormous reduction of the amount of input data and some form- and position invariance at the top. These local invariances are essential for succeeding parallel-sequential pattern recognition mechanisms [7]. Consequently the **MIP** can be considered as a set of separate, pyramidal organized topographical maps of the visual scene. Each of these analyzing maps includes at every resolution level parameter filters for different elementary features (texture density, colour and intensity contrasts, spatial frequency). In this way any single input location is split into a multiple parametrical description at several resolution levels. By weighted superposition of the neural activity between the several feature maps an encoding of high syntactic complexity (many different feature detectors activated at the same time) into an blurred activity distribution at the top of the pyramid is realized (Fig. 2 - right side) . The more such different parameter filters are triggered by a certain visual location, the stronger the total activation of the corresponding area in the highest pyramid level will be. Up to now feature maps for local contrasts in the intensity and for differences in the colour (hue) distribution have been implemented in **MIP** (see Fig. 1). The importance of other pathways will be analyzed in psychophysical eye movement experiments in future. In the contrast analyzing pathway (Fig. 1) for each node (i,j) at all resolution levels of this pyramid the following non-linear contrast selection based on a Laplace-filtering in the W-domain of the colour space is performed.

$$y_{ij} = \hat{y}\left(1 - \exp\left(-\frac{d_{ij}^2}{(\alpha\tilde{s})^2}\right)\right) \tag{1}$$

with

$$\tilde{s}^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2$$

Each d_{ij} is the result of a local Laplace-filtering, α determines the shape and $\tilde{s}$ the turning-point (threshold) of the transfer function, $\hat{y}$ is the maximum of intensity.

In the colour processing pathway (Fig. 1) for each node (i,j) the Euclidean distance to the average hue in the $M_{rg}M_{yb}$ plane is computed and finally scaled with the corresponding intensity W_{ij}. The weighted superposition of the top level activities of the various feature extracting pathways is a critical point of our model concept since no detailed experimental data are available about this. Therefore we are not able to specify the weights α, β, γ of the feature map superposition (see Fig. 1) exactly. Only estimated parameters providing plausible simulation results can be proposed. In the context of the psychophysical experiments mentioned above these aspects of weighted superposition have to be analyzed too.

2.3 The Selection and Control Network

The **Selection and Control Network - SCN** realizes a kind of cortically controlled selection of the most conspicuous locations in the visual field which have been encoded as peaks within the activity landscape of the **MIP** superposition plane. When the input to **SCN** has various activity peaks because of several striking locations in the visual scene (Fig. 2-right side), the network is to select not simply the maximum one but successively that peaks with the highest competition energy in the landscape (extension and altitude of the activity bubble). Then in result of internal relaxation processes in this network and of a controlled top-down manipulation of the lower processing levels the most conspicuous locations of the visual scene will be activated one after another (Fig. 3). Since the first relaxation process toward a stationary solution needs some time, a higher **knowledge based processing level** has enough time to activate its acquired knowledge about the presented activity landscape as a whole, about spatial aspects of the landscape composition and about the individual

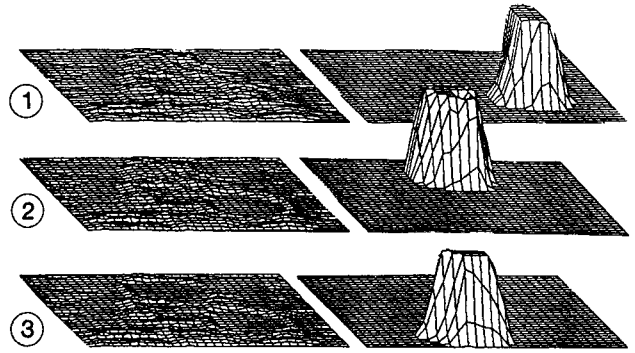

Figure 3: *Activity dynamics in Selection and Control Network at the first processing cycles (1-3). (Left side) shows several input activity distributions top-down manipulated by SCN during the preattentive search. (Right side) shows the sequential development of single SCN-decisions on the manipulated activity landscapes of the left side (see 3.).*

form of the bubbles. By a topographically correct feedback from that higher processing level to **SCN** (see Fig. 1) this activated internal knowledge is able to modulate directly the activity dynamics of the relaxation and sharpening process in **SCN** [7]. By this superposition of preattentive and attentive search dynamics in the same subsystem a continuous control of the input data stream is possible. Without internal knowledge about the activity landscape only the preattentive search determines the dynamics in **SCN**, otherwise complex interactions between both processes will occur.

The Selection and Control Subsystem has been implemented as controllable single-layer neural network connected feedback. The activity dynamics of each neuron of this system can be described mathematically by the following differential equation:

$$T_1\frac{dy_{ij}(t)}{dt} + y_{ij}(t) = \Phi\left(\epsilon x_{ij}(t) + c_{ij}(t) + \mu\sum_{\substack{k=i-a \\ l=j-a}}^{\substack{i+a \\ j+a}}(w_{ijkl}y_{kl}(t)) - \nu I(t)\right) \tag{2}$$

with the nonlinearity

$$\Phi\left(z_{ij}(t)\right) = \begin{cases} 0 & : \quad z_{ij}(t) < 0 \\ z_{ij}(t) & : \quad \text{else} \end{cases} \tag{3}$$

$\epsilon x_{ij}(t)$ denotes the weighted input to each SCN-neuron (i,j), c_{ij} stands for the top-down control signal from the higher knowledge based processing levels to each neuron (i,j), μ and ν denote the coupling parameters for cooperative and competitive interactions within the network, T_1 is the time constant of the controlled network. The w_{ijkl} are the components of the synaptic weight matrix **W** coding a Gaussian shaped filter kernel that realizes the desired local cooperation between the SCN-neurons. The global inhibition $I(t)$ is controlled by a proportional-integral controller with the following integral equation:

$$I(t) = (\hat{y} - \tilde{y}(t)) + \frac{1}{T_n}\int_0^t (\hat{y} - \tilde{y}(t))\, dt \tag{4}$$

$\hat{y}$ stands for the control point of the subsystem (maximum output activity of **SCN**) while $\tilde{y}(t)$ results from the maximum of the current output activities of all $y_{ij}(t)$. The time constant of the PI-controller (T_n) and the coupling parameter ν of the controlled variable $I(t)$ are determined in the z-plain because of the discrete-time simulation. Both parameters are adjusted corresponding to the parameters ϵ , μ and T_1 of the network and to the wanted characteristics of this subsystem.

3. Dynamics of the Model and Concluding Remarks

After presenting an input to the hierarchy distinct decisions on the input are developing in **MIP** and **SCN** because of the different syntactic complexity at the several locations in the input (see Fig. 3). The local cooperation and ensemble competition between the neurons in the **SCN** suppress activity peaks with weaker competitive power and sharpen the remaining one so that only that peak with the highest energy (location with largest local input complexity) survives (see Fig. 3/1 - on the right). The required shifting of the selective attention is performed by a **SCN**-feedback controlled manipulation of the channel transfer characteristics in the several pathways and processing levels of **MIP**. These channel-specific control mechanism take the selected decision for a certain time out of discussion. In this way the next decision can develop only on the remaining parts of the input and the next-grade complex input configuration will start this search process anew (Fig. 3/1-on the left). In a time sharing manner other activity peaks of the landscape (coding related input segments of different conspicuity) can be selected, creating a time-sharing sequence of internal decisions. In this way our system decomposes a complex visual input into a sequence of striking local input segments arranged according to its local syntactic complexity (see Fig. 3/2, 3/3). Such a parallel in sequence decomposition of a complex input scene is a fundamental mechanism for self-organization of sequential and episodical representations, a special type of non-trivial dynamical knowledge, that is acquired in the higher knowledge processing levels of our model concept.

Without a knowledge based top-down manipulation of the relaxation dynamics in SCN, the established sequence depends only on the local conspicuity of the different input locations. The attentive selection based on internal systems knowledge about spatial relations and form features of the activity landscape has not been discussed in this paper but it can be realized by the knowledge controlled modulation of the preattentive dynamics mentioned above. Therefore the adaptive neural network architecture **GNOM** operating as a parallel-sequential link between data driven pattern analysis and knowledge controlled attentive search has been developed and prepared for implementation in the hierarchy [7]. This coupling is an object of research at present.

References

[1] **Koerner,E., Tsuda,I., Shimizu,H.** Parallel in Sequence–Towards the Architecture of an Cortical Processor. In Parallel Algorithms and Architectures, Akad.-Verl. Berlin 1987, 37-47

[2] **Koerner, E., Boehme, H.-J.** Organization of an Episodic Knowledge Data Base. In Proceedings of ICANN91, vol. 1, pp.873-878, North-Holland 1991

[3] **Koch,C., Ullmann, S.** Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(1985) p. 219-227

[4] **Orban,G.A.** Neural operations in the visual cortex. Springer Bln., Hdbg., NY, Tokyo 1984

[5] **Giefing, G.-J., Janßen, H., Mallot, H.-P.** A Saccadic Camera Movement System for Object Recognition. In Proceedings of ICANN91, vol. 1, pp.63-68, North-Holland 1991

[6] **Dow, B.M.** Colour Vision. In Vision and Visual Dysfunction, Vol. 4, The Neural Basis of Visual Function, (Ed.) G. Leventhal, pp. 316-338, The Macmillan Press, 1991

[7] **Gross,H.-M., Koerner, E., Pomierski,T.** GNOM—A Modular Network Architecture for Adaptive Parallel-Sequential Pattern Recognition. Proc. of ICANN91, vol. 1, 747-752, North-Holland 1991