Proc. of 39th Int. Scient. Coll. Ilmenau, 1994, pp. 124-131

A Neural Network Architecture for Self-Organization of Object Understanding*

D. Heinke, H.-M. Gross

Technical University of Ilmenau, Division of Neuroinformatics
98684 Ilmenau, Germany
e-mail: dietmar@informatik.tu-ilmenau.de

1 Introduction

This work is based on experimental findings in psychophysics, which show that a visual scene is processed in parallel and in sequence. This findings led to the theory of visual attention[15] [13]. Following the searchlight metaphor we consider a visual attention process as an internal scanning, which is accomplished over a visual scene without any eye movements. This internal scanning build up the frame for our hypothesis on a behaviour-oriented self-organization of object understanding.
Biological information processing aims at the generation of adequate behaviour in relation to incoming sensor signals of the environment. Thus, the biological system develops behaviour-oriented information processing [11] [6]. In our understanding the internal scanning is just another behaviour like gripping an object. We propose that one objective of the internal scanning could be to achieve a temporal representation of objects. In other words parts of an object are bind together in temporal nearness by means of the scanning process. Thus, the behaviour leads to a solution of the binding problem [3].
The main goal of our work is to investigate how this kind of object understanding could be achieved in a self-organizing manner by a neural network architecture. In our terms self-organization implies, besides that there exists no teacher but the environment, no explicit different system states between learning and execution. Thus, we consider self-organization as an on-line learning process.
In order to simplify the problem we restrict the environment to 2D-real-world-scenes. Objects of these scenes mainly vary with respect to translation and have slight distortions concerning rotation and luminance. Thus, the neural network architecture needs to have translation invariant processing abilities and cope those distortions.
This paper describes the essentials of the whole neural network architecture.

2 Neural Network Architecture

2.1 Basic Assumptions

*supported by BMFT, Grant No. 413-5839-01 IN 101D - NAMOS-Project

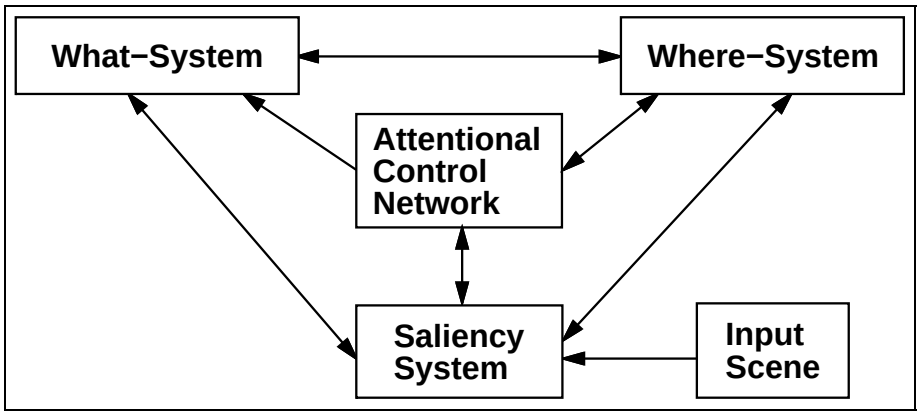

Figure 1: *This Figure gives an overview of the neural network architecture. It is derived from biological and psychophysical models (see text).*

There are three basic assumptions which led to the neural network architecture (Fig. 1). The first assumption is the correctness of the psychological models of selective attention [13] [15]. These models claim that there exist a parallel and a serial stage of processing visual scenes. The parallel stage computes a saliency measure for every location of a scene. The measure seems to depend on the whole scene. On which features it relies is still a subject of ongoing discussion. If a search target is somewhat more complex, a serial stage performs a search across a scene. This search seems to be guided by the parallel stage [15]. In our model we assume, that the behaviour of the serial stage is similar to the external scanning [8], in the sense that internal scanning might be a scanning without a "go-signal" [4].
The second assumption follows the division into a *What-System* and a *Where-System* of the visual system [14]. The what-pathway leads through V4 into Inferotemporal Cortex and the where-pathway into the Posterior Parietal areas. The Pulvinar might act as an *Attentional Control Network* for the whole system. A detailed discussion on that topic is given in [6].
The third assumption concerns the self-organization using 2D-real-world-scenes, whose objects essentially vary with respect to translation. In this environment objects consist of a set of stable spatial relations between features. This property can be used by a self-organization of object understanding. This motivates the need of a structure of the neural network architecture which is capable of learning that set of stable relations between features. The neural network architecture presented here is only capable of learning stable spatial relations. How to combine this result to a set of relations associated with an object will be discussed in the final section and in [6] [2].

2.2 Overview of the Behaviour

Following the third assumption of the previous section a neural network architecture that aims at a self-organization of object understanding should acquire knowledge about the stability of spatial relation of features and apply this knowledge in order to control the scanning process. Hence the neural network architecture has the following behaviour: The *Attentional Control Network* selects the first location of the focus of attention using the output of the *Saliency System*. According to the contents of the focus the top-down-

control (*What-* and *Where-System*) selects a set of stable moves to certain features. The top-down-control searches through this set and looks for features and locations of the input scene that matches with one element of the set. If it finds such a *successful move*, this move will be the starting point for the next search. If no *successful move* can be found, the *Attentional Control Network* will select a move on the base of the *Saliency System*. Hence a *data-driven move* is made. After a while the *Attentional Control Network* enforces a *data-driven move*, a so called *forced move*, using a certain criterion (Sec. 2.6). This effect aims at leaving one object and move to another object. Because top-down-control learns all the time, its behaviour improves during search as well as from *data-driven moves*.

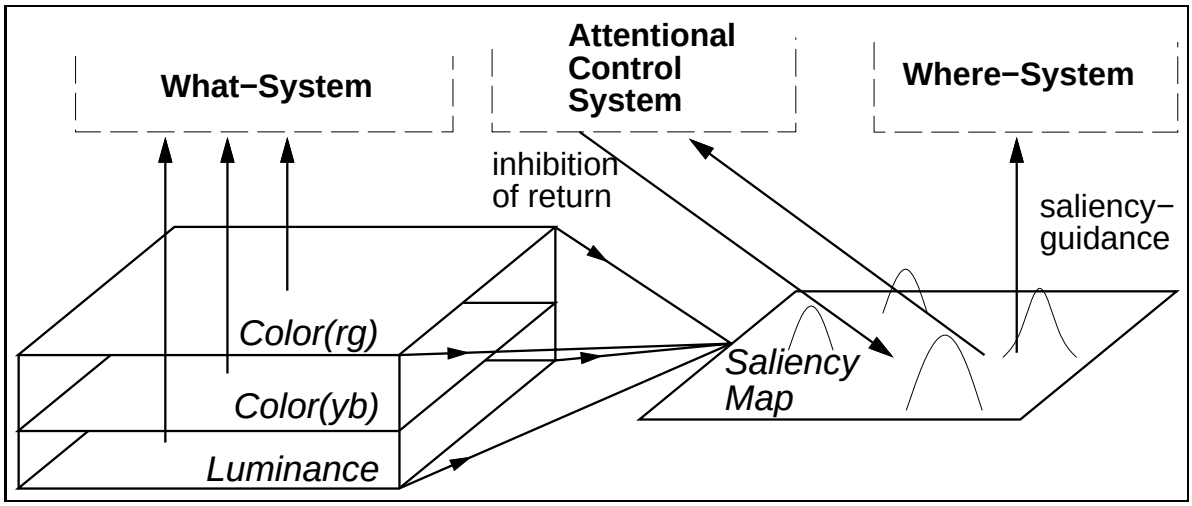

Figure 2: *The Saliency System computes a scene-dependent measure of saliency of scene locations from the features luminance and difference color space.*

2.3 Saliency System

The *Saliency System* (Fig. 2) consists of a feature extraction and a computation of the *Saliency Map*. The first part computes three feature maps from input image: luminance map, red-green map and yellow-blue map [7]. They are used for computation of the saliency measure. Following psychophysical experiments saliency is a scene-dependent measure. We used two sorts of measures leading to two maps and than we combined these maps by a weighted sum to the *Saliency Map*: The first map uses a soft-threshold function applied to DOG-filtered luminance image, whose threshold is scene-dependent. The second map computes the Euclidean distance to the average hue for each location of the scene [7].
In addition, the three maps are an input of the *What-System*. In order to reduce the amount of data and to be robust against distortions a multiresolution pyramid was used for every map.

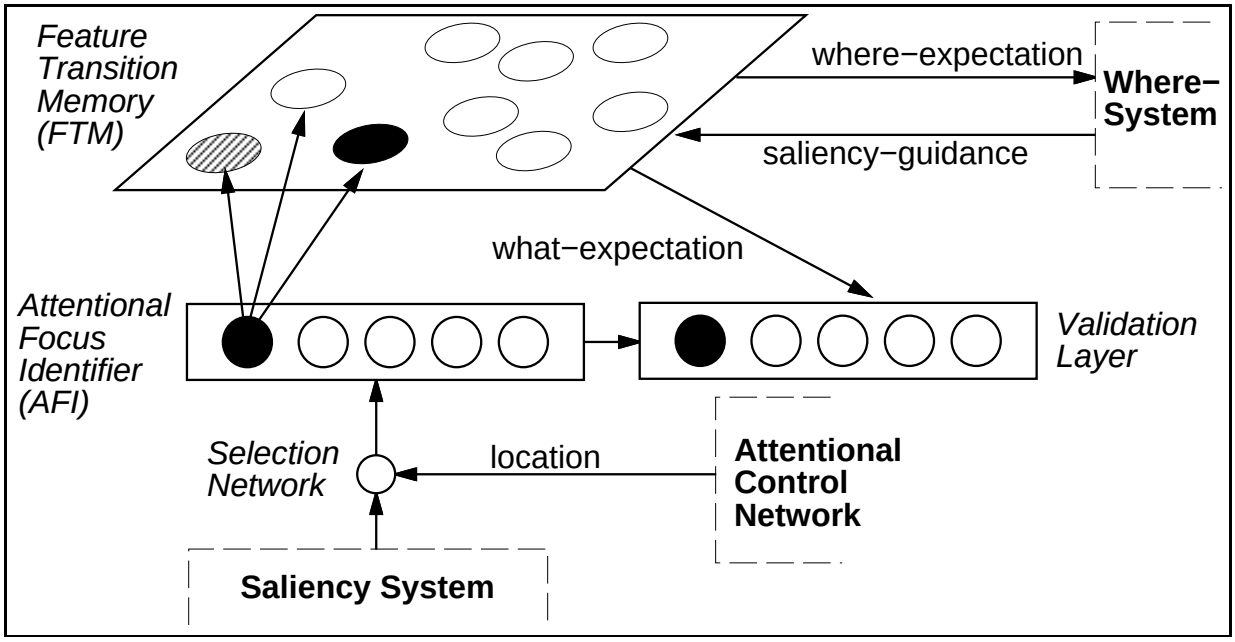

Figure 3: *The What-System is the main part of the neural network architecture. It learns stable moves between features through weight-values of the FTM.*

2.4 What-System

Fig. 3 depicts the details of the *What-System*. The functionality of this system incorporates the knowledge of stable spatial relation between features. The *Selection Network* extracts the focus of attention out of the feature maps of the *Saliency System* by sigma-pi nodes which perform a correlation between location information of the *Attentional Control Network* and the feature maps. This approach is similar to [12]. The result is fed into the *Attentional Focus Identifier* (AFI) which is simply a Kohonen feature map [10]. The output of the AFI leads to an activation of nodes of the *Feature Transition Memory* (FTM) which is modulated by the *Saliency Map* that is mapped through the *Where-System*. The weights between AFI and FTM give a measure for the stability of a move to certain features. A *Winner-Take-All* (WTA) mechanism selects the node with the highest activity and generates a what-expectation and a where-expectation. Hence the WTA selects a move to a location that might be stable and salient. The *Validation Layer* compares the features of the resulting location with the expectation. If it matches, a *successful move* is found. Otherwise the WTA generates another expectation and the search continues.

2.5 Where-System

The *Where-System* (Fig. 4) has two functions: First, the *Location Map* combines feature based information of the *What-System* with spatial information. Second, the *Position Transformation* maps relative spatial information to absolute spatial information and vice versa. Thus, the *Position Transformation* enables the neural network architecture to translation invariant processing. For coding the spatial information a 2-dimensional

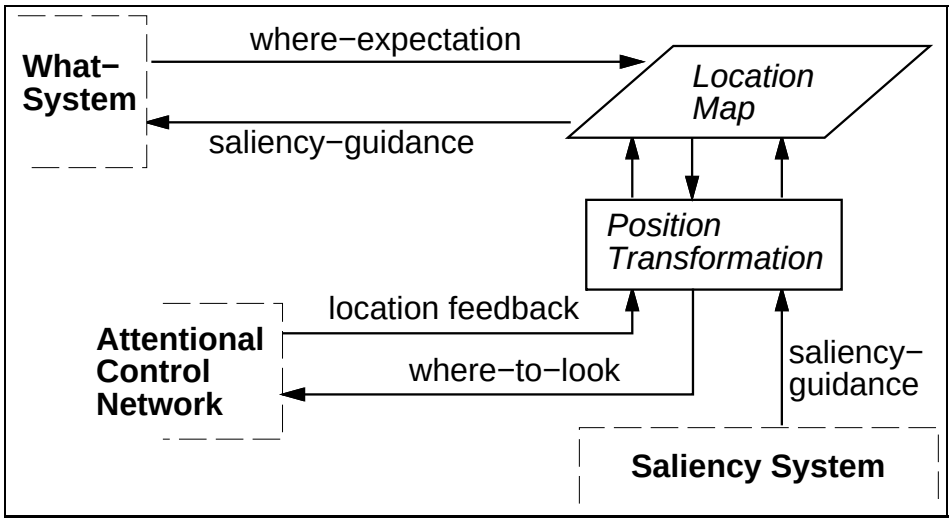

Figure 4: *The Where-System maps the activity of the What-System into spatial information through the Location Map and vice versa. The Position Transformation transforms relative position coding into absolute position coding and vice versa.*

activity distribution is used. This leads to a homogenous way of position coding within the whole system. Sigma-pi-units, that perform convolution or correlation, can transform absolute position coding into relative position coding and vice verse [9].

2.6 Attentional Control Network

Main goal of the *Attentional Control Network* (Fig. 5) is to determine which system, the top-down-control or the *Saliency Map*, is allowed to select the next location of the focus of attention. A WTA-network generates an activity peak at the most salient location of the scene. This activity is sent to the *What-System*, if there is no activity from the *Where-System* otherwise the *Where-System* controls the location of the focus of attention. The purpose of the *Inhibition Map* is to suppress a return to a location of a *successful move*. The *Internal Inhibition Map* stores every location where a *data-driven move* is made to, no matter if there is a *successful move* or not. This map plays an important role during learning (Sec. 3). The feedback to the *Where-System* tells where has been moved to indeed.
If the sum of activity of the *Saliency Map* drops under a certain threshold a *forced move* is made and the activity of the internal inhibition map is removed. This effect aims at leaving one object and move to another object.

3 Self-Organization Process

The main issue of self-organization is to maintain an on-line learning and find out about stable spatial relations of features in order to achieve an object understanding. This is done by simple reinforcement learning approach [1]. The matching signal of the *Validation Layer* is used as reinforcement signal to update the weights of the FTM. Thus a stable

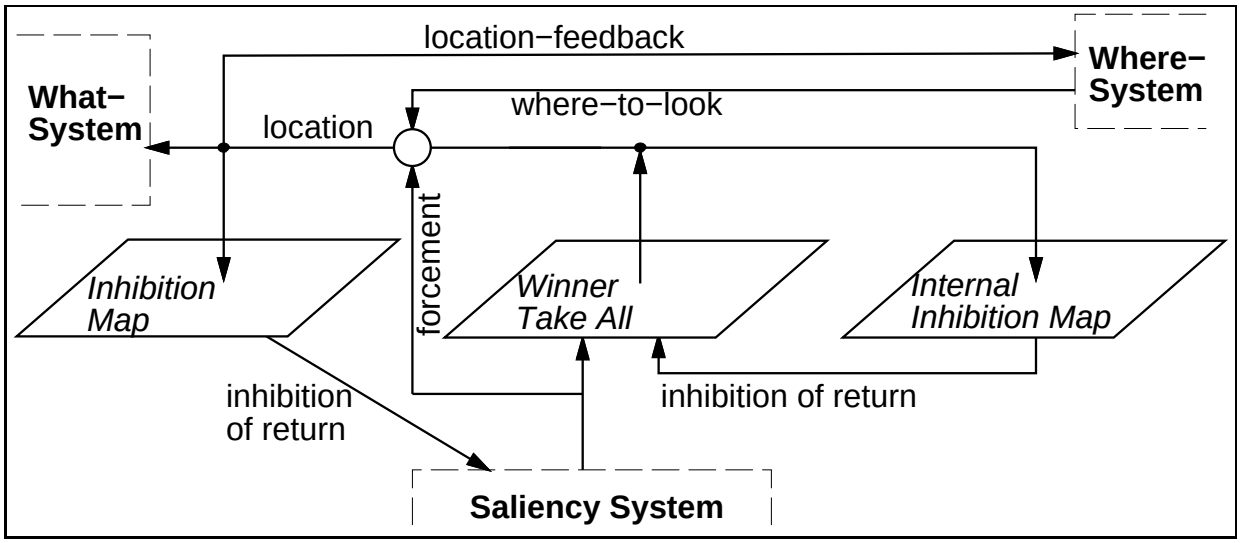

Figure 5: *The Attentional Control Network determines which system, top-down-control or Saliency System, is to select the next focus of attention. In addition it inhibits the return to locations already processed.*

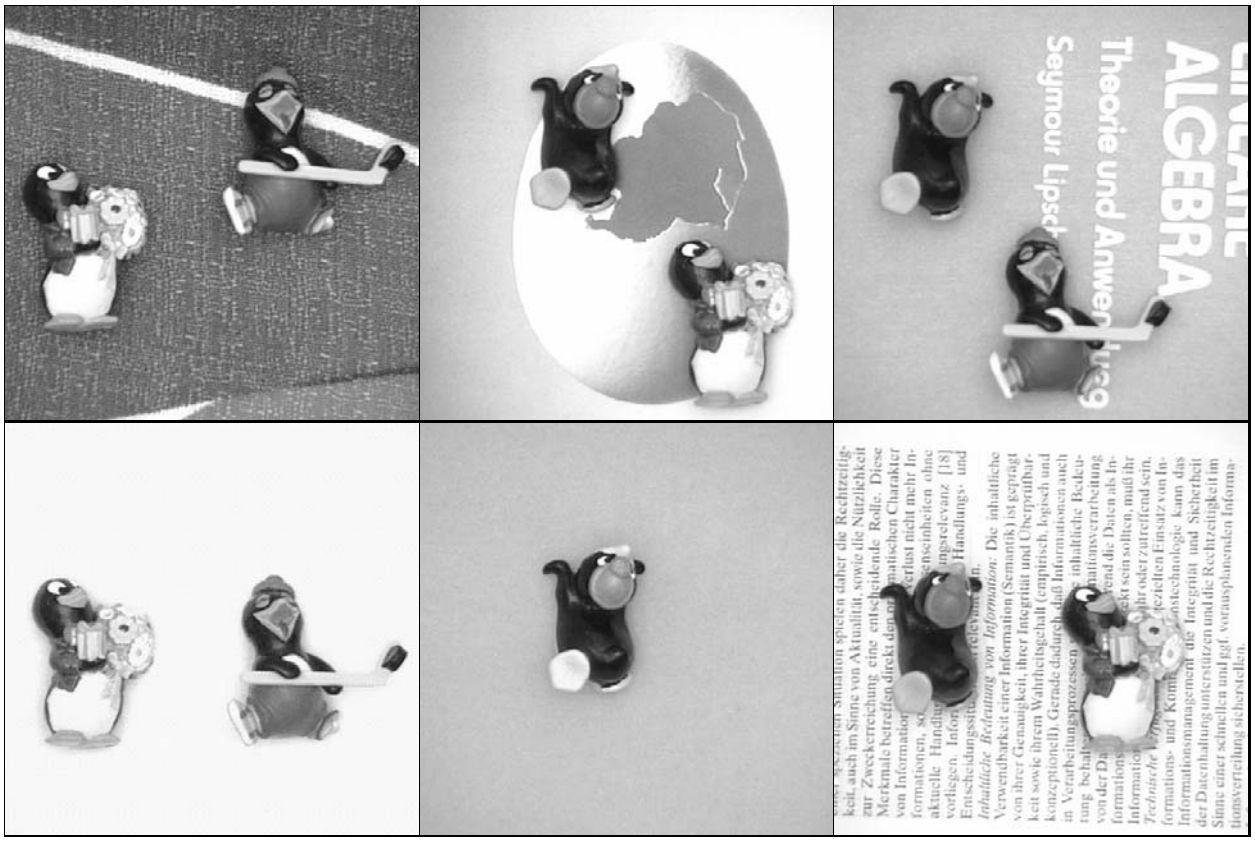

Figure 6: *Examples out of 100 color scenes presented to the neural network architecture.*

move to certain features gains much reinforcement whereas unstable spatial relations of features gain less reinforcement. In addition the weights between *Where-* and *What-System* are also modified in order to generate a correct expectation.
In order to have a successful learning strategy the FTM performs two different learning phases for every move: a passive and an active phase. During the passive phase the FTM only tests the move to corresponding features but does not execute the move. When the move is tested frequently enough, the FTM switches to the active phase. During the active phase the move is always executed if it is successful. There are two reasons for introducing a passive phase: First, it speeds up learning because on every focus of attention the FTM can learn about the stability of all moves as long as they are all in the passive state. Second, a control of the scanning process can be only successful on the base of sound knowledge, otherwise the control would move into some kind of local minimum. The passive phase might contradict psychological data, due to the inherent time limits. On the other hand, this strategy might correspond to what happens if a human being sees millions of scenes a day.
The self-organization process discussed so far achieves that the scanning stays more and more on one object rather than switching between objects. Because of this behaviour a second object gets more and more attractive and a *data-driven move* would very likely lead to another object. Thus, the top-down-control has very little chance to learn more about the actual object. To avoid this trap the *Internal Inhibition Map* was introduced (see Sec. 2.6).

4 Results

Fig. 6 shows examples of scenes presented to the neural network architecture. The images of Figure 7 illustrate the learning progress achieved by the self-organizing process. They show that the top-down-control successively taking over the control of the scanning process. Because the right object is more salient than the left one the control for the right object starts earlier than for the left object. Due to noise and to different spatial relations of different objects the way of controlling the scanning process changes during learning process. Finally, the control of the scanning process achieves temporal binding of parts of the objects and the *forced move* switches between the objects. The object understanding is not perfect, because the last move of the left object (lower right image) leads to the background. This results from a lack of complete object understanding, which is achieved by learning the set of stable spatial relations associated with an object.

5 Conclusion

We introduced a neural network architecture that is capable of a simple self-organizing and on-line learning of an object understanding. The approach is based on extraction of stable spatial relation between features. By means of real-world-scenes we demonstrated how the learning process works. But the result also show that the object understanding is not complete. In order to get a complete object understanding it is necessary to learn about sets of stable relations of features. This will be achieved by combining the approach presented here with the episodic knowledge-base introduced in [2].

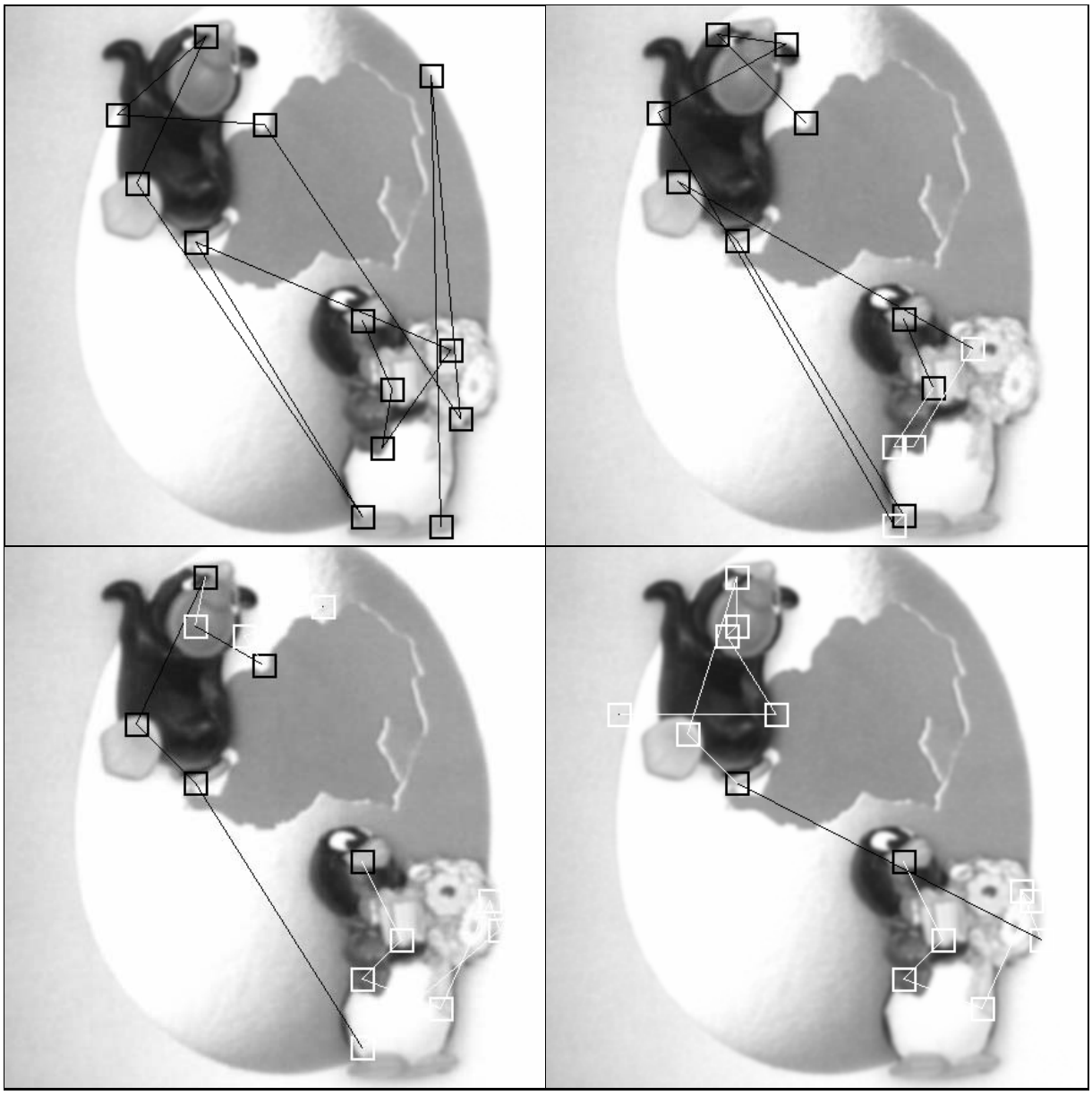

Figure 7: *These figures illustrate the learning process of the neural network architecture. In order to get a better presentation the images are enlarged. The white boxes show a focus of attention reached via a successful move and the black boxes show the result of a data-driven move. The upper left picture depicts a data-driven scanning process. The following images document the progress of top-down-control after 20, 50 and 100 scenes. Thus, the top-down-control increases its influence on the scanning process leading to an object understanding.*

Following the latest results in psychophysics the shape of the focus of attention is determined by the Gestalt laws [5]. Thus it is necessary to enhance the shape of the focus of attention towards a compact region of attention.

References

[1] Barto,A. G., Sutton,R. S., and Anderson,C. W. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *Reprinted from IEEE Transactions on Systems, Man, and Cybernetics*, 13(5):835–846, 1983.

[2] Boehme,H.-J., Braumann,U.-D., and Gross,H.-M. A Neural Network Architecture for Episodic Feature Representation. *this volume*, 1994.

[3] Crick,F. and Koch,C. Towards a neurobiological theory of consciousness. *The Neurosciences*, 2:263–275, 1990.

[4] Desimone,R., Wessinger,M., Thomas,L., and Schneider,W. Attentional Control of Visual Perception: Cortical and Subcortical Mechanisms, 1990.

[5] Enns,J.T. and Rensink,R.A. Preattentive Recovery of Three-Dimensional Orientation From Line Drawings. *Psychological Review*, 98(3):335–351, 1991.

[6] Gross,H.-M., Boehme,H.-J., Heinke,D., and Pomierski,T. An Behaviour-Oriented Approach to an "Object-Understanding" in Visual Attention. *this volume*, 1994.

[7] Gross,H.-M., Franke,R., Boehme,H.-J., and Beck,C. A Neural Network Hierarchy for Data Driven and Knowledge Controlled Selective Visual Attention. In *14. DAGM-Symposium*, 1992.

[8] Hacisalihzade,S. S., Stark,L. W., and Allen,J. S. Visual Perception and Sequences of Eye Movement Fixations: A Stochastic Modeling Approach. *IEEE Trans. on System, Man, and Cybernetics*, 22(3):474–481, 1992.

[9] Heinke,D. and Gross,H.-M. A Simple Selforganizing Neural Network Archiecture for Selective Visual Attention. *ICANN-93*, pages 63–66, 1993.

[10] Kohonen,Teuvo. Self-organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43:59–69, 1982.

[11] Mallot,H.A., Kopecz,J., and Seelen,W.v. Neuroinformatik als empirische Wissenschaft. *Kognitionswissenschaft*, 3(1):12–23, 1992.

[12] Olshausen,B., Anderson,C., and Essen,van D. A Neural Model of Visual Attention and Invariant Pattern Recognition. *CNS Memo 18*, 1993.

[13] Treisman,A.M. and Gelade,G. A Feature-Integration Theory of Attention. *Cognitive Psychology*, 12:97–136, 1980.

[14] Van Essen,D.C., Anderson,C.H., and Felleman,D.J. Information processing in the primate visual system: An integrated systems perspective. *Science*, 255:419–423, 1992.

[15] Wolfe,Jeremy M. and Cave,Kyle R. Deploying Visual Attention: The Guided Search Model. *AI and the Eye*, pages 79–103, 1990.