in: Proc. of ICNN'95, IEEE Int. Conf. on Neural Networks 1995, Perth, Australia, pp. 657-662, IEEE Press

A Behaviour-oriented Approach to an Implicit "Object-understanding" in Visual Attention*

H.-M. Gross, D. Heinke, H.-J. Boehme, U.-D. Braumann, T. Pomierski
Technical University of Ilmenau, Dept. of Neuroinformatics
98684 Ilmenau (Thuringia), Germany
e-mail: homi@informatik.tu-ilmenau.de

ABSTRACT

We present a hypothesis and a neurobiologically motivated neural architecture for self-organization of a behaviour-oriented, implicit "object-understanding" in the context of an attention based scene analysis. The paper emphasizes the functional architecture for self-organization of an "object-understanding" on the basis of internal anticipation, reshuffling and evaluation of the scanning process. Finally, by means of chromatic real-world scenes we demonstrate the effect of an emerging "object-understanding" on the scanning process - evident as drastical modification of the spatio-temporal scanning behaviour.

1. Selective visual attention and "Object-understanding"

Selective attention is a widely accepted mechanism explaining the decomposition of a fovealized complex visual scene into a sequence of reliably detectable input components. Numerous publications on visual attention, for instance [13], [1] or [3], emphasize the purpose of visual attention to focus the limited neural resources for recognition on specific regions within this scene. Also, in our behaviour-oriented approach to visual perception presented here, this data-driven and knowledge controlled dissolution of the highly parallel visual input into meaningful components, which can be reassembled in a flexible way to complex perceptual structures, is of fundamental importance. Especially the knowledge controlled decomposition is a prerequisite for handling unknown scenes or objects. Therefore, of our particular interest are cortical learning and control principles, that facilitate a selective manipulation of the scanning dynamics in the course of scene analysis.

Most of the known approaches are oriented on modeling the preattentive search or the biologically motivated top-down control, but especially the problem of unsupervised on-line learning - as a prerequisite for a flexible control - is usually not considered. Therefore, we present a neural architecture able to generate and verify hypotheses on the further progress of the internal scanning process *(Sensory Controlled Internal Simulation)* and to learn actively considering the conformity between the internally anticipated and the real data-driven scanning process. In this context, anticipation means the generation and testing of hypotheses about which meaningful components (**what**) are to be expected **when** and **where** in the visual field – this is an internal simulation of a real scanning process. In our understanding, this internal scanning of a fovealized visual scene is expression of a behaviour similar to the external eye-movements during saccadic scene analysis.

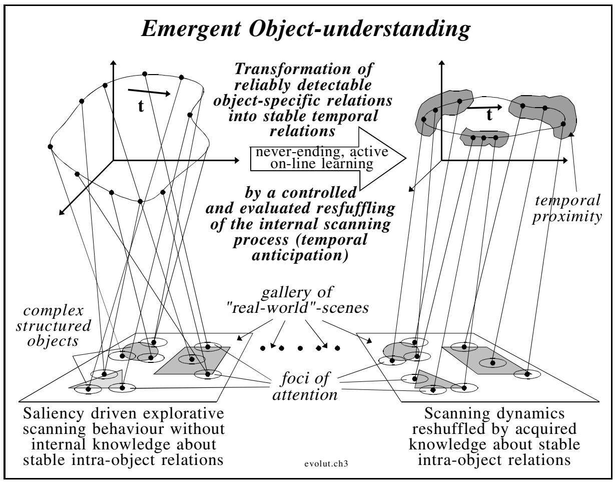

Fig. 1: *Evolution of stable temporal relations in attentional vision as expression of extractable stable intra-object relations, to be interpreted as an emerging behaviour-oriented understanding which components of the visual scene belong together. Based on this acquired implicit object-understanding, the proposed neural architecture is able to manipulate the data-driven scanning behaviour more and more to make it more effectively and to accelerate the internal decision processes.*

In our concept, self-organization of an "object-

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

understanding" means, that typical, reliably detectable striking visual components and their object-specific spatial relations detected during preceding scannings more frequently, gradually can be learned and coupled in the temporal domain (see Figure 1). This way, an *active reshuffling in time* of the scanning process is achieved. This is necessary to bring those input components into *temporal proximity*, which make some sense together but are not yet proper coded in the spatio-temporal data stream. We postulate, that the evolution of temporal proximity in scanning behaviour is to interpret as a simple behaviour-oriented 'understanding' which components of a visual entity (object) belong together. In our opinion, temporal proximity in attentional processing could be a well-suited, possibly the only criterion for an autonomous segmentation and learning of unknown objects arranged in highly structured visual scenes. In this sense, our proposed model is to demonstrate the on-line evolution of a characteristic scanning behaviour selecting the relevant input components belonging to the same object *successively in time* - without the need of a preceding explicit training of all relevant objects in a special learning mode.

2. Functional Architecture of the Attentional Model

Based on the known facts from neurophysiology, neuroanatomy and psychophysics (for more details see Figures 2 and 3), we developed a neurobiologically inspired model able to generate internal attentional focus movements as significant systems behaviour. This model is to decompose a fovealized retinal image into a sequence of *reliably detectable components* ranked by its visual conspicuousness and controlled by the already acquired knowledge. To establish a consistent internal representation, the attentional search has to be guided by the already acquired knowledge from the beginning. Additionally, the data-driven search dynamics should be *reproducable* as good as possible under varying conditions (illumination, scene composition, etc.). Therefore, we introduced several robust adaptation mechanisms in the 'early-vision' colour processing levels of our model (see [11]).

Our neural architecture is composed of several interacting subsystems which define basic abilities and information processing tasks **a)** to yield a measure of the conspicuity of locations within a complex structured scene, **b)** to select the most salient regions of the scene in a topographic organized Saliency Map, **c)** to shift the focus of attention from the current to the next striking location and **d)** to anticipate the scanning dynamics and to control the preattentive flow of information taking into account

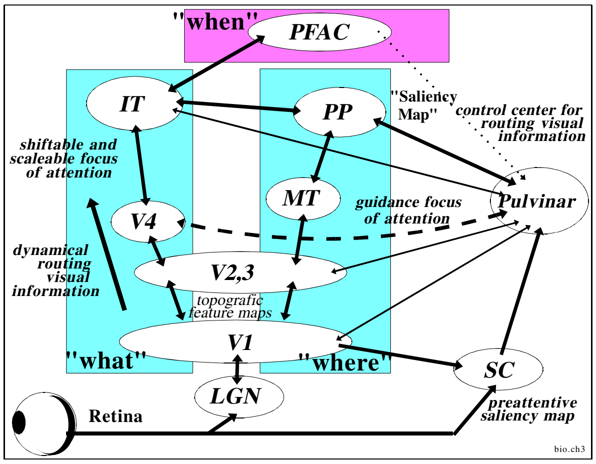

Fig. 2: *Major visual processing pathways of the primate brain that have been considered in our model. Information from the retino-geniculo-striate pathway enters the visual cortex through area V1 and then proceeds through a hierarchy of visual areas that can be subdivided into two major functional pathways. The so-called* **"what"***-pathway leads through V4 and Inferotemporal Cortex (IT) and is mainly concerned with object-feature identification, regardless of position or size. The* **"where"***-pathway leads into the Posterior Parietal areas (PP), and seems to be concerned with the locations and spatial relationships among objects, regardless of their identity. As proposed in* OLSHAUSEN *[10], we consider the PP as a "saliency map" representing the locations of potential attentional targets in the scene. The Pulvinar may play an important role in providing the control signals required for dynamical routing and modulating the information flow from V1 to IT. Referring back to [9], the Prefrontal Association Cortex (PFAC) is considered as highest organizational level for learning, planning and dynamical control of the temporal explorative behaviour in our concept (***"when"-system***).*

the already acquired knowledge about stable intra-object relations.

Figure 4 shows a simplified scheme of the model architecture and the *main processing levels*. Loci of spatio-temporal feature discontinuities (striking regions within the fovealized scene) are detected in parallel by a *data-driven feature analysis*. Based on a dynamic routing of these located retinal information, an *attentional identification process* analyzes the selected details of the scene. This routing process from retina to cortex is called *internal scanning* and is consistent with the "searchlight metaphor" proposed by TREISMAN [13] and ANDERSON [1]. For the routing, we implemented a simplified version of the OLSHAUSEN-ANDERSON model [10]. As OLSHAUSENS's approach, our model belongs to the so-called "input-gating" class of neural models of attention. The key action of this attention mechanism is to route selectively 'interesting' regions of the visual scene onto higher 'cortical' processing levels.

The identification of the actual focus of attention

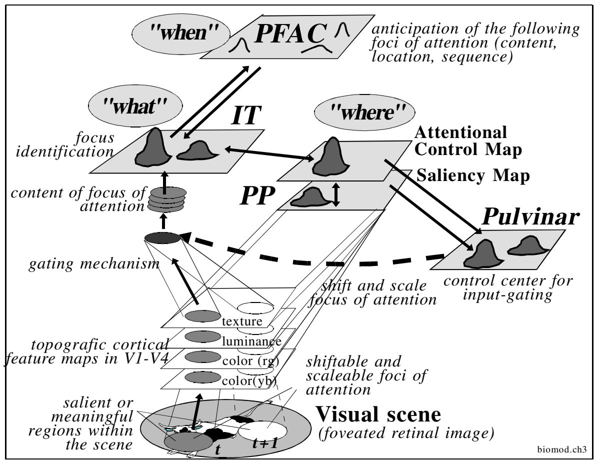

Fig. 3: *Translation the relevant neurobiological facts from Fig. 2 into a principal functional architecture, that underlies our model in Fig. 4.*

(*'what'*) and its spatial relations to the previous focus ('where') are stored in two separate memory sections of our *Feature Transition Memory* ('where-what' - separation.) The following *Episodic Object Memory* integrates the attentional shifts extracted and predicted successfully by the *Feature Transition Memory* and tries to establish and to verify internal object hypotheses by claiming shifts to scene locations specific for that objects (top-down control). The attentional and the data-driven processing pathways feed their target demands into the *Attentional Control Map* representing the whole scene in parallel. This map decides 'when and where' to shift the focus of attention. A detailed discussion of the lower subsystems (see Fig. 4) is given in [7].

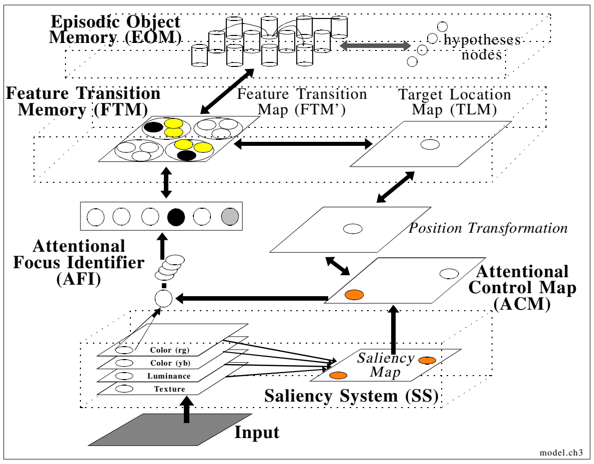

Fig. 4: *Simplified functional architecture of our attentional model.*

2.1. Saliency System (SS)

The reliable detection of striking regions within the input is a prerequisite for the evolution of an 'object-understanding' during explorative scanning the scene. Therefore, we developed a *Saliency System* [4], that has been influenced essentially by the neurophysiological concepts of primary visual processing in a variety of maps for different elementary features, such as texture, contrast, colour or motion. [8]. In our model, we utilize a very simple measure of saliency based on "luminance-texture-colour" pop-out (see Fig. 5). The goal of the *Saliency System* is the reliable detection of differences in local conspicuous features of the input in separate analyzing pathways and their encoding in an activity landscape within a *Saliency Map*. The state of each of these maps therefore signals how conspicuous a given location in the visual scene is. By weighted superposition of the neural activity in the different feature maps, an encoding of high syntactic complexity (many different feature detectors activated at the same place and the same time) into a blurred activity distribution in the *Saliency Map* is realized.

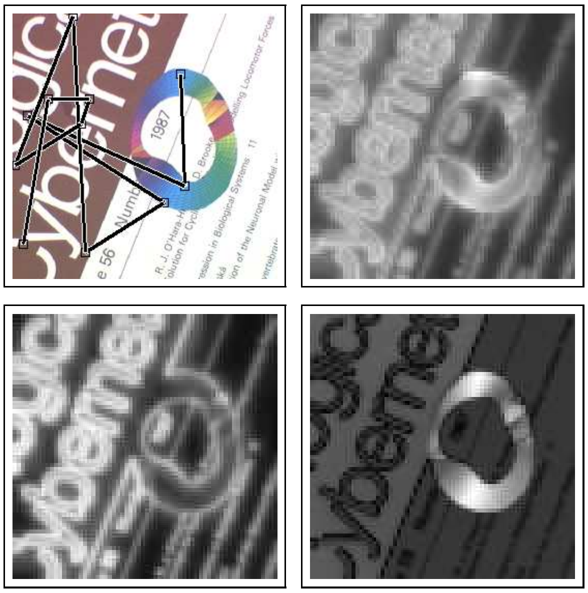

Fig. 5: *Computation of a Saliency Map of a typical input scene (top left) as prerequisite both for the preattentive and the attentive scanning process. For that purpose, differences in local and global conspicuous features are detected in distinct analyzing pathways. Up to now pyramidal organized feature maps for local contrasts in the intensity (bottom left) and for global contrasts in a new, neurobiologically inspired colour space (bottom right) (see [11]) have been implemented in our Saliency System. By weighted superposition of these activity maps an encoding of saliency or high syntactic complexity (see text) in the Saliency Map (top right) is realized. This superposition is a critical point since no detailed experimental data are available about this. Only estimated parameters providing plausible simulation results can be proposed. The marked scan-path (top left) shows the course of a data-driven preattentive search and the sequence of selected most striking input locations.*

2.2. Attention Control Map (ACM)

The objective of this control map is to guide the focus of attention to salient or meaningful regions of the visual input. Therefore, the ACM carries out a sequential search for the most striking locations within the visual field which have been encoded as peaks within the activity landscape of the ACM. When the input to the map has various activity peaks because of several salient locations in the visual scene, the network is to select not simply the maximum one but successively the peaks with the highest competition energy in the landscape. This way, the *Attention Control Map* generates a sequence of decisions controlling control the foci of attention to route their contents to the *Attentional Focus Identifier* and the *Feature Transition Memory*. The ACM is modulated both bottom-up by the *Saliency System* and top-down by spatio-temporal expectations from the *Feature Transition Memory*. This way, the map and the routed sensory data are controlled by activated hypotheses (*What items - where in the visual field ?*), so that the interesting components can be reshuffled in time according to the state of internal hypothesis activation. The higher subsystems communicating with ACM (see Fig. 4) register their activity maps only after an appropriate *position transformation* into ACM and vice versa. For that purpose, the absolute position coding within the *Attention Control Map* is transformed into a relative position coding of distance and direction between subsequently following foci of attention (more see [7]).

2.3. Attentional Focus Identifier (AFI)

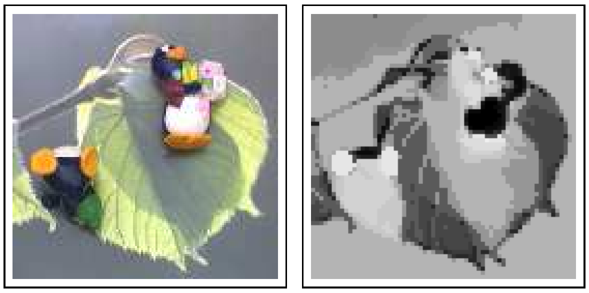

Fig. 6: *Processing properties of the Attentional Focus Identifier demonstrated in context of a segmentation problem of a typical chromatic scene from the gallery (left). Local input regions having similar local feature sets (colors, texture) show very similar local classification results (right) because of the topographic organization of the Neurons within AFI. The local classification results are shown as grey-values in a 64*64 grid.*

The *Attentional Focus Identifier* (AFI) operates on the focus of attention controlled by the *Attentional Control Map*. It determines the similarity of the actual attentional focus feature set to the feature sets extracted and learned in previous scanning cycles. In this context, we implemented and tested different unsupervised learning neural networks (Self-organizing Feature Maps (SOFM), Neural Gas, FuzzyART). Because of their topology preserving properties, the SOFM show very robust and good reproducable classification results (see Fig. 6) - a prerequisite for the extraction of stable and specific intra-object relations during the analysis of image sequences.
In the *Validation Layer* of the *Target Focus Identifier*, a substructure of the AFI (see Fig. 7), the feature sets of the foci predicted by the *Feature Transition Memory* as *'what-where-expectations'* are verified and evaluated returning reinforcement signals to the system. This task is called *specific hypothesis verification* and is fundamental for the active learning in the *Feature Transition Memory*.

2.4. Feature Transition Memory (FTM)

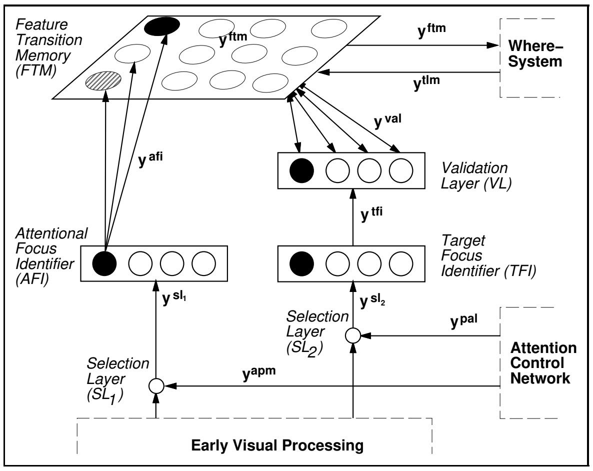

Fig. 7: *More detailed information flow within the "what"-system of our model between the different neural subsystems for focus identification and selection of the next shift in the scan process ("what-where"-expectation).*

In the context of our behaviour-oriented approach an object can be described as a temporal scanning sequence of meaningful components belonging together. Therefore, FTM extracts and learns continuously stable intra-object transitions using the statistics of the explorative behaviour during preceding scanning processes. Preattentive inter-object shifts can not be stabilized sufficiently in this memory, since the objects in different scenes usually vary in their spatial arrangements (see Fig. 8). The FTM learns stable feature-position relations between subsequent foci of attention by linking position codings, that occur repeatedly in the *Target Location Map*, with corresponding feature transitions in the *Feature Transition Map (FTM')* (see Fig. 4 and 7). The feature transitions are learned by mo-

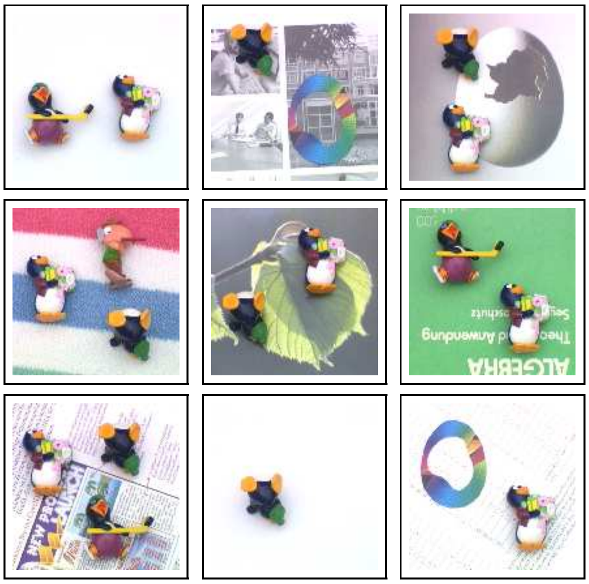

Fig. 8: *Some examples out of the gallery of typical 'real-world' scenes composed of structured objects (not learned explicitly and therefore initially unknown to the system) and varying complex background situations for simulating the self-organization of a behaviour-oriented implicit "object-understanding". All objects show relatively stable intra-object relations between salient or meaningful components. Since the objects vary in the scenes with respect to translation, illumination and view, we get unstable inter-object and object-background relations.*

difying the weights between the AFI and the FTM', providing a measure for the stability of a move from one region to the next expected region. An internal *Validation Layer* (see Fig. 7) continuously evaluates the success of a 'feature-position' prediction and controls the learning process actively by giving reinforcement signals to the *Feature Transition Map* and the *Target Location Map*. So, a reproducable transition from one interesting region to another certain region gains much reinforcement, whereas unstable 'what-where-relations' gain less reinforcement.

The comparatively unspecific and very local hypotheses generated by this subsystem should rather be considered as a preliminary stage for an implicit object-understanding since more global temporal relationships between the movements cannot be extracted and represented yet. This was the crucial motivation for the development of a second attentional level, the *Episodic Object Memory*.

2.5. Episodic Object Memory (EOM)

This subsystem is the highest organizational level of our architecture, where both internal anticipation and dynamical control of the following scanning process take place (see Fig. 4). While in the FTM only unspecific hypotheses on the next most likely

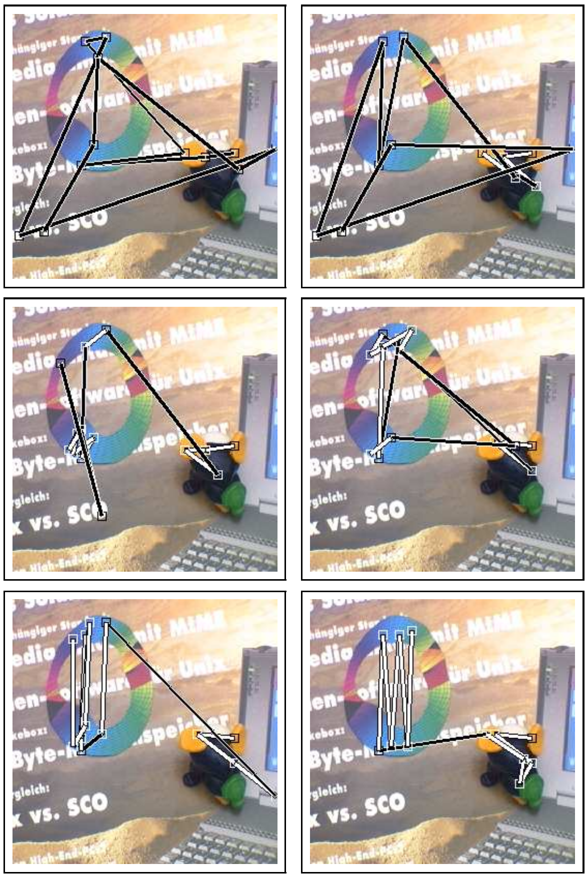

Fig. 9: *Temporal evolution of an emergent object-understanding to be seen as modification of the scanning behaviour of the same scene after presenting 20, 30, 40, 60 and 90 other scenes out of the gallery (from top left to bottom right). (top left) The marked shifting of the focus of attention shows the course of the data-driven search and the decomposition of the scene into a sequence of striking local input components arranged according to their local conspicuousness. The sequence of the transitions is shown by superimposed black lines. Without any internal knowledge about typical intra-object relations ("what-where-when") the system is not able to establish suitable hypothesis about objects to anticipate an effective scanning. In result, numerous shifts occur between the penguin and the salient structures in background. The system is not able to select the striking components of the same object successively in time. (bottom right) Result of the evolution process - this figure illustrates a completely knowledge controlled scanning process starting from the penguin. The white lines connect those foci of attention that have been driven successfully by the EOM/FTM on the base of the self-organized 'what-where-when' object-knowledge acquired during the last 90 scenes.*

move can be generated and verified, the *Episodic Object Memory* tries to take longer sequences of successfully predicted focus transitions and to keep them as candidates for whole objects or parts of objects. For that purpose, the components of different scanning sequences are mapped into characteristic memory traces within the columnar organized *Episodic Object Memory*. So, each sequence of decisions on certain input components is transferred into a spatio-temporal representation within the *EOM*, to be activated for a specific top-down control. The *Episodic Object Memory* interacts reciprocally with the *Feature Transition Memory* in so-called 'hypothesize-verification-cycles' and tries to control the course of the attentional search by generating more global 'what-where-expectations'. Via this feed-back the memory is able to search for such input components which support best one of the activated object-hypotheses. A more detailed description of the neural architecture of the *Episodic Object Memory* is presented in [2].

3. Simulation Results

Fig. 10: *Another instructive example for a drastical modification in the scanning behaviour in result of the evolution of an implicit 'object-understanding' about different visual structures (penguins).*

In Fig. 8 we present a gallery of typical scenes for self-organization of a behaviour-oriented "object-understanding" in attention based scene analysis. In these scenes different unknown objects (penguins) are arranged randomly in various locations. Also numerous different background situations are used to achieve unstable inter-object relations. During presenting this scene gallery, our model has been able to extract and to learn *stable intra-object relations* ("what-where-when") autonomously and to evolve an implicit object-understanding. Compared with the original data-driven search dynamics in Fig. 9 and 10 (top left) a drastical modification in the scanning behaviour can be observed in these figures. In conformity with Fig. 1, shifts between different objects are reduced heavily, now all striking components of the same object are selected successively in time. This much more effective explorative behaviour is an expression of that emerging *implicit object-understanding.* It is the result of transforming reliably *detectable object-specific relations* uncoupled with respect to time at the beginning into more and more *stable temporal relations* by active learning and temporal reshuffling the scanning process. A detailed description of the model architecture, the several neural subsystems, the activation dynamics within these subsystems and the reinforcement based active learning mechanisms is presented in [6].

References

[1] **Anderson, C.H. & Van Essen, D.C. (1987).** Shifter Circuits: A computational strategy for dynamic aspects of visual processing. *Proc. of the Nat. Acad. of Science*, 84, 6297-6301

[2] **Braumann, U.-D.,Boehme, H.-J, Gross, H.-M. (1995).** An Episodic Knowledge Base for Object-Understanding. *Proc. of ESANN'95, Brussels*, pp. 175-180

[3] **Desimone, R. (1992).** Neural Circuits for Visual Attention in the Primate Brain. In: *Neural Networks for Vision and Image Processing*, 343-364, Cambridge, Mass.: MIT Press.

[4] **Gross, H.-M., Koerner, E., Boehme, H.-J. (1992).** A Neural Network Hierarchy for Data and Knowledge Controlled Selective Visual Attention. *Proc. of ICANN'92, Brighton*, Vol.2, pp. 825-828, Elsevier Science Publisher

[5] **Gross, H.-M., Boehme, H.J., Heinke, D., Pomierski, T. (1994)** Self-Organizing a Behaviour-Oriented Interpretation of Objects in Active-Vision. *Proc. of ICANN '94, Sorrento*, Vol. 1, pp. 58-61, Springer

[6] **Gross, H.-M., Boehme, H.J., Heinke, D., Pomierski, T. (1995)** Fundamental Cognitive Mechanisms in Neural Architecture. *Final NAMOS-project report*, German Federal Dept. of Education and Research (BMBF), (in German)

[7] **Heinke, D. & Gross, H.-M. (1994).** A Neural Network Architecture for Selforganization of Object Understanding. In: *Proc. of Int. Scient. Coll'94, Ilmenau*, pp. 124-130.

[8] **Koch,C. & Ullmann, S. (1985).** Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*,4, 219-27

[9] **Kupfermann, I. (1991).** Localization of Higher Cognitive and Affective Functions: The Association Cortices. In: Candel, Schwartz, Jessell (eds.) *Principles of Neural Science*, Appleton & Lange, pp. 823-838

[10] **Olshausen, B.; Andersen, C. & van Essen, D. (1993)** A Neural Biological Model of Visual Attention and Invariant Pattern Recognition. *Journal of Neuroscience.* vol. 13, 4700-4719

[11] **Pomierski, T. & Gross, H.-M. (1995).** A Neurobiological Approach for Perceptual Colour Modification. *Proc. of DAGM'95, Bielefeld*, Springer, 1995 (in German).

[12] **Robinson, D.L. & J. McClurkin (1989)** The visual superior colliculus and pulvinar. In: *The Neurobiology of Saccadic Eye Movements*, pp. 337-358, Elsevier Science Publisher

[13] **Treisman, A. (1983).** The role of attention in object perception. In Braddick, O.J., Sleigh, A.C. (Eds.), *Physical and Biological Processing of Images*, Springer, 1983