to be published in **Proc. of 39th Intern. Scientific Colloquium Ilmenau**, Sept. 27 to 30, 1994

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

H.-M. Gross, D. Heinke, H.-J. Boehme, T. Pomierski

Technical University of Ilmenau, Dept. of Neuroinformatics
98684 Ilmenau (Thuringia), Germany
email: homi@informatik.tu-ilmenau.de

1 Behaviour-oriented "Object-understanding" ?

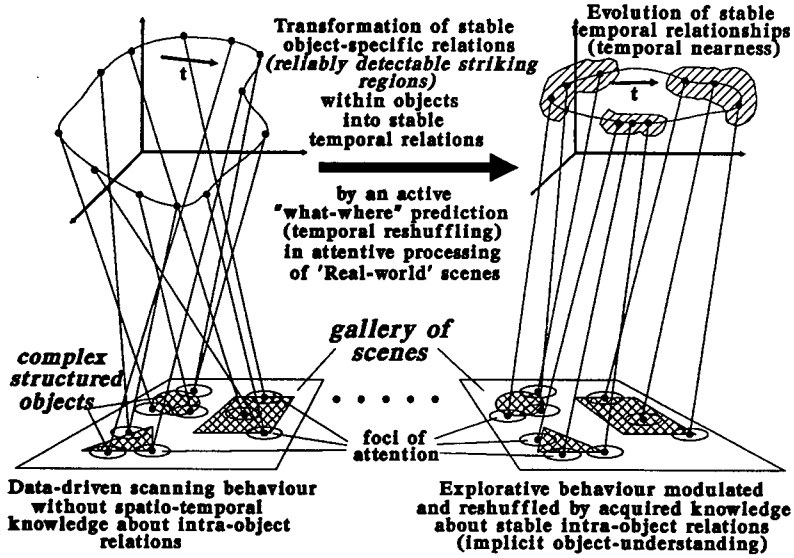

Figure 1: *Evolution of stable temporal relations in attentional vision as expression of stable object-specific relations within the input, to be interpreted as a behaviour-oriented understanding which components of the visual scene belong together within the same object. Based on the searchlight-metaphor, the proposed functional architecture generates explorative internal attentional focus movements* **(left)**. *This sequential scanning is the base for self-organization of an implicit "object-understanding" by controlled temporal reshuffling the data-driven exploration process* **(right)**.

In this paper we present a biologically plausible hypothesis and a neural architecture for self-organization of a *behaviour-oriented 'object-understanding'* in the context of an attention based scene analysis. The focus is on the functional architecture and the dynamical principles suited for self-organization of knowledge about complex visual structures, and for a *behaviour-oriented "understanding"* of real-world objects expressed in a typical explorative behaviour in scanning the scene (see Fig. 1). The basic idea of our visual attention concept is that the visual input to a real-world analyzing system is not a set of preselected, figure-ground segregated objects which are to be properly arranged, but the system itself

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

has to select reliably detectable input components out of the massively parallel input (see Fig. 2) (Gross, 1992). *Selective attention* is a widely accepted mechanism explaining this decomposition of a complex visual scene into a sequence of salient components.

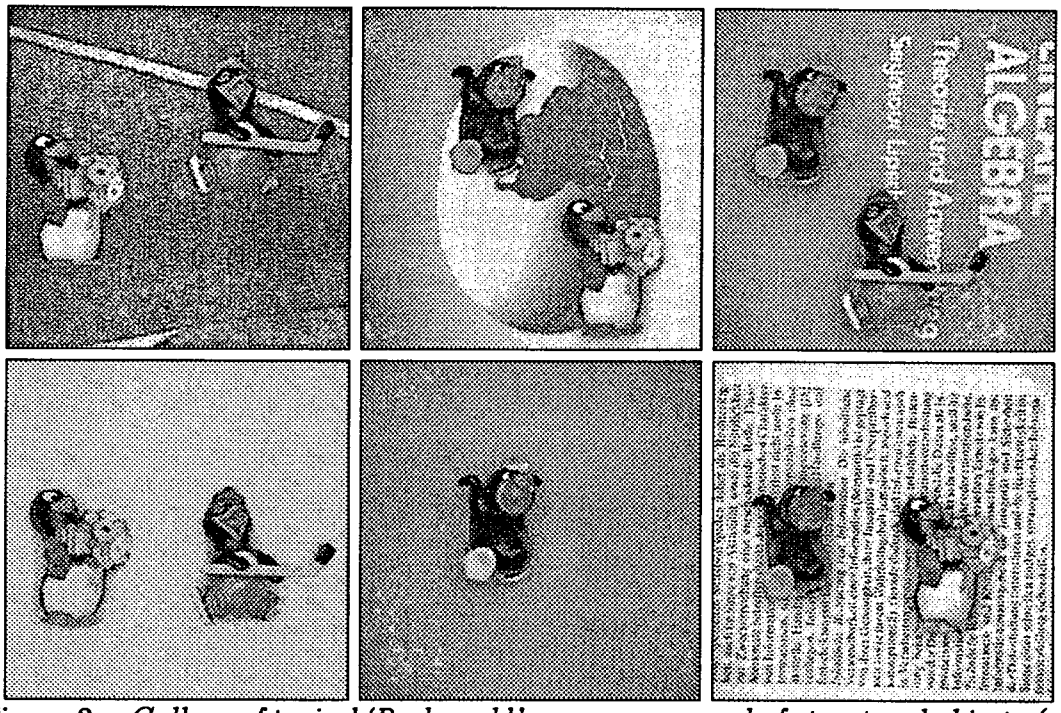

Figure 2: *Gallery of typical 'Real-world' scenes composed of structured objects (not learned explicitly and therefore unknown to the system) and varying complex background situations for simulating the data driven preattentive search and the self-organization of a behaviour-oriented implicit 'object-understanding'. All objects show relatively stable intra-object relations between salient or meaningful components, respectively between their feature sets (colour, texture, luminance). Since the objects vary in the scene with respect to translation, illumination and view, we get unstable inter-object and object-background relations. These instabilities within the different scenes are the base for self-organizing an 'object-understanding' in our concept.*

Numerous publications on visual attention, for instance (Treisman, 1983; Anderson, 1987; Desimone, 1992), emphasize the purpose of attention to focus the limited neural resources for recognition on a specific region within the scene. This spatial area can vary in size and position and scans objects in the visual field at a rate of about 30-50 msec per location. Therefore, an important aspect of our approach is the data and/or hypotheses-driven dissolution of the highly parallel visual input into meaningful components, which can be reassembled in a flexible way to new complex visual structures. This decomposition is a prerequisite for handling *unknown scenes* or objects. Of our particular interest are such concepts like generation and active verification of hypotheses about the input in a feedback coupled internal perception process, the so-called *Sensory Controlled Internal Simulation.* In the context of an explorative systems behaviour, that means the generation and testing of hypotheses about *what* components are to be expected *when* and *where* in the visual field – this is an internal anticipation of a real spatio-temporal visual attention process. In our understanding, this internal scanning of a complex scene is a behaviour similar to the external eye-movements in saccadic scene analysis or to an autonomous motion in the external world. In our concept, self-organization of an "object-understanding" means, that typical, reliably detectable striking visual components and their object-specific spatial and temporal

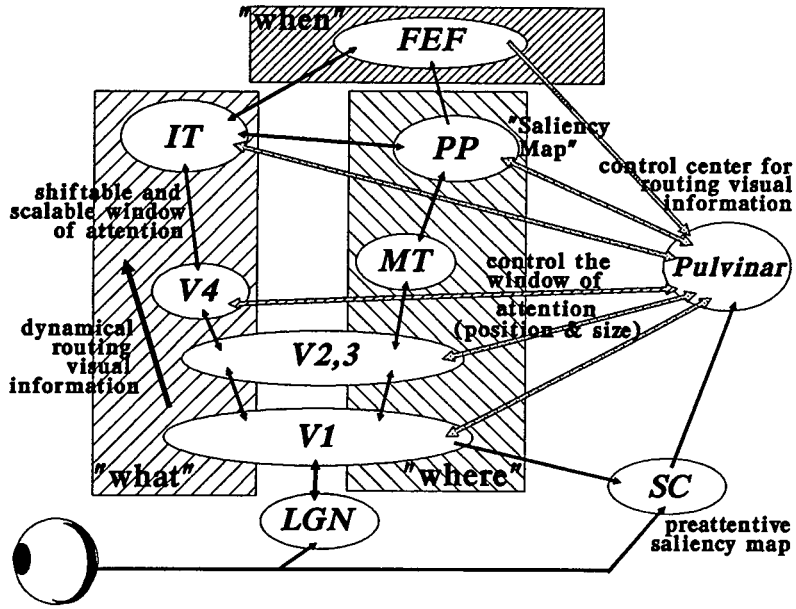

Figure 3: *Major visual processing pathways of the primate brain that have been considered in our model. Information from the retino-geniculo-striate pathway enters the visual cortex through area* **V1** *and then proceeds through a hierarchy of visual areas that can be subdivided into two major functional 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. The* **pulvinar***, a sub-cortical nucleus of the thalamus, makes reciprocal connections with all these cortical areas (Robinson 1992). As proposed in (Olshausen 1993), 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. The frontal eye fields (***FEF***) may act as highest organizational level for learning, planning and dynamical control of explorative behaviour.*

relations detected in preceding scannings more frequently, gradually can be coupled or linked in the temporal domain (see Figure 1). This way, an *active reshuffling in time* of the striking visual components can be achieved. This is necessary to bring those input components, which make some sense together but are not yet proper coded in the spatio-temporal data stream into *temporal nearness*. We postulate, that the evolution of stable temporal relations (temporal nearness) in scanning behaviour is an expression of stable object-specific relations within the input that is to interpret as a behaviour-oriented 'understanding' which components of a visual entity (object) belong together. In our hypothesis, *temporal nearness in attentional processing*, could be a well-suited, possibly the only criterion for an unsupervised segmentation and learning of unknown objects arranged in highly structured visual scenes. In this sense, our attentional model is to demonstrate the self-organization of a characteristic scanning behaviour, to select the striking input components belonging to the *same object* successively in time. To our mind, an explorative behaviour like this is an adequate expression of an acquired *implicit object-understanding* - without the need of explicit training the relevant objects in a special learning mode before.

2 Functional Architecture of the Attentional Model

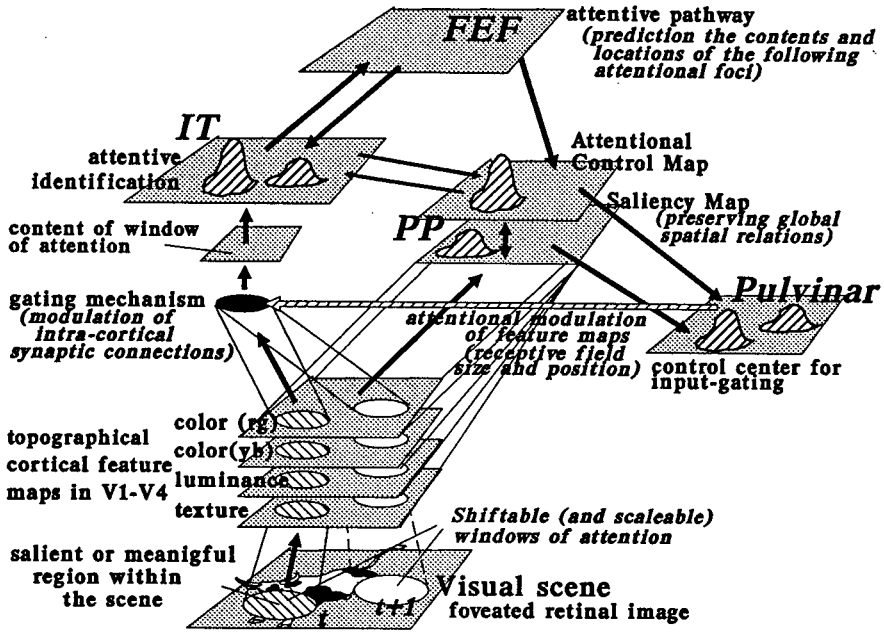

Figure 4: *Translation the relevant neurobiological facts from Fig. 3 in a principal functional architecture, that underlies our model in Fig. 5.*

Based on the known facts from neurophysiology, neuroanatomy and psychophysics (for more details see Figures 3 and 4), we developed a neurobiologically inspired model that is able to generate internal attentional focus movements as expression of a systems behaviour. This model is to decompose a complex visual input into a reverberating 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 based from the beginning on the already acquired knowledge. Additionally, the data-driven (preattentive) search dynamics has to be *reproducable* as good as possible. Therefore, we implemented expensive, 'intelligent' and robust feature extracting mechanisms in the 'early-vision' levels (see Pomierski, 1994). Our model 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 and to select the most salient regions of the scene in a topographic Saliency Map, b) to shift the focus of attention from the current to the next striking location and c) to control the preattentive flow of information in the course of attentional processing taking into account self-organized knowledge about stable intra-object relations.

Figure 5 shows a simplified scheme of the model architecture and the *main processing levels.* Loci of spatio-temporal feature discontinuities (striking regions of the scene) are detected in parallel by a *data-driven feature analysis.* Based on a dynamic routing 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 (1983) and Anderson (1987). For the routing we implemented a simplified version of the Olshausen-Anderson model (1993). The intention of that model is to provide a neurobiologically plausible mechanism for shifting and rescaling the representation of an object from its retinal reference frame into an object

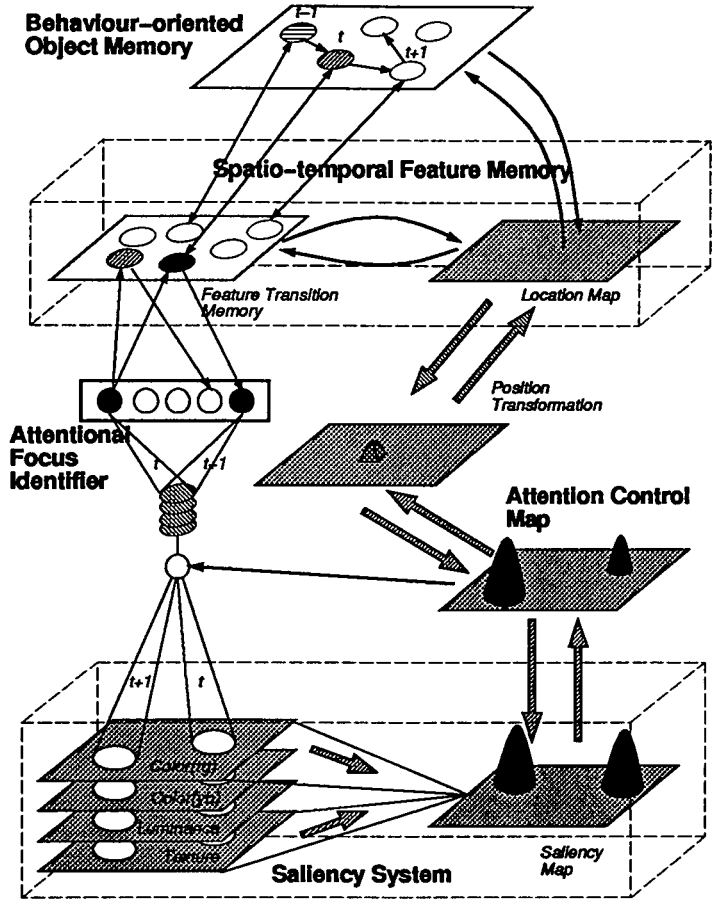

Figure 5: *Overview on our attentional model (more explanations see text).*

centered reference frame in higher cortical areas. As Olshausens's approach, our model belongs to the so-called "input-gating" class of neural models of attention (following the classification suggested by Desimone, 1992). Here, the key action of attention is to route selectively 'interesting' regions of the visual scene onto higher 'cortical' processing levels. The behaviour-oriented 'object-representation' is established in our model implicitly between several communicating subsystems. So, the identification of the actual focus of attention *('what')* and its spatial relations to previous foci ('where') are stored in two separate memory sections of our *Spatio-temporal Feature Memory* ('where-what' - separation.) The following *Behaviour-oriented Object Memory* integrates the attentional shifts predicted successfully by this *Feature Memory* and tries to verify object hypotheses by claiming shifts to scene locations specific for that object (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. 5) is given in (Heinke, 1994 and Pomierski, 1994). In (Boehme, 1994), some fundamental investigations about a special type of a neural architecture suited for a *Behaviour Oriented Object Memory* are presented.

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

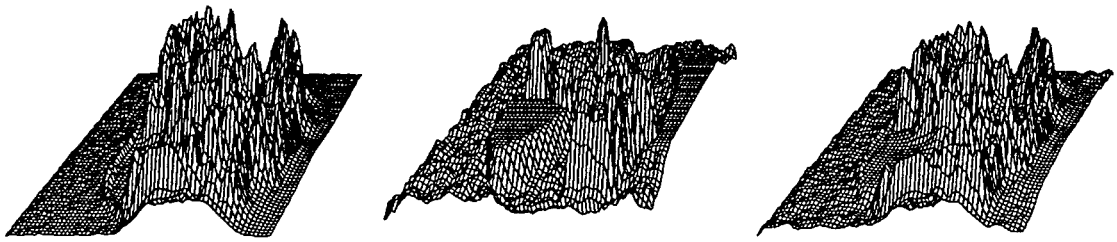

Figure 6: *Different topographic feature maps of the scene shown in Fig. 7 based on a transformation of the primary colour space (RGB) into a new neurophysiologically motivated $WM_{rg}M_{yb}$-colour space derived from the red/green and yellow/blue colour opponent system of LMS-neurons (DeValois, 1993). The purpose of the multi-parametrical description of the scene in feature maps is the detection of differences in local conspicuous features of the input and their encoding in a Saliency Map* **(right)**. *Therefore feature maps for local intensity contrasts* **(left)** *and for differences in the colour distribution* **(middle part)** *have been implemented in our Saliency System (Gross 1992, Pomierski 1994). For instance, in the colour processing pathway (in the middle) for each neuron (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}. By weighted superposition of the different feature maps the intended encoding of local input complexity in the Saliency Map (right) is realized. Since no detailed experimental data are available about this, we can use only estimated parameters providing plausible simulation results.*

Saliency System (Gross, 1992), 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. (Koch & Ullmann, 1985). In our model, we utilize a very simple measure of saliency based on "luminance-texture-colour" pop-out. 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*. By this analysis, those locations that differ significantly from their surrounding are singled out at each processing level. 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 between the implemented feature maps, an encoding of high syntactic complexity (many different feature detectors activated at the same place at the same time) into an blurred activity distribution in the *Saliency Map* is realized (see Fig. 6).

2.2 Attention Control Map (ACM)

The objective of this control map is to guide the attentional window 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 striking locations in the visual scene (see Fig. 6), the network is to select not simply the maximum one but successively that peaks with the highest competition energy in the landscape. This way, the *Attention Control Map* generates a sequence of decisions which control the 'foci of attention' to route their contents to the *Attentional Focus Identifier* and the *Spatio-temporal Feature Memory*. The ACM is modulated both bottom-up by the *Saliency System* and top-down by spatio-temporal expectations from *Spatio-temporal Feature Memory*. This way, this map and the routed sensory data can be 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. When a local area has been identified, the window of attention has to be shifted to another interesting part of the scene. Therefore, when a group of ACM-neurons is active for some time (long enough for recognition to take place), they begin to shut off by self-inhibition through a delay. This allows other blobs within the map to compete successfully for control of the window of attention. Based on this self-inhibition, the attention switches to the next most conspicuous location. Additionally, we implemented a further inhibitory section of this map ('inhibition of return') to prevent a return to already focused positions. Nevertheless, a relaxational term was introduced in both inhibitory sections of the map which allows to foveate locations after some time again. The higher subsystems communicating with ACM (see Fig. 5) 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 Heinke, 1993).

2.3 Attentional Focus Identifier (AFI)

The *Attentional Focus Identifier* 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 autonomously in previous scanning cycles. The implemented network architecture has a certain similarity to an ART-architecture (unsupervised learning, on-line learning, using a distance metric). Moreover, in the AFI the feature sets of the foci predicted by the *Spatio-temporal Feature Memory* as 'what-where-expectations' are verified giving reinforcement signals to that system. This task is called *specific hypothesis verification* and is fundamental for the autonomous learning in the *Spatio-temporal Feature Memory*.

2.4 Spatio-Temporal Feature Memory (STFM)

In the context of our behaviour-oriented approach an object is considered as a temporal sequence of meaningful object components belonging together. Therefore, the *Spatio-temporal Feature Memory* is learning continuously stable intra-object relations using the statistics of the explorative behaviour during preceding search 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. 2). The self-organization is a result of interactions between *Attentional Focus Identifier*, *Feature Transition Memory (FTM)* and the *Location Memory (LM)* (see Figure 5). Based on these interactions, the *Spatio-temporal Feature Memory* learns stable feature-position relations between subsequent foci of attention. That means, the *LM* links stable position codings which occur repeatedly with corresponding feature transitions in the *FTM*. The feature transitions are learned by modifying the weights between the AFI and the FTM, giving a measure for the stability of a move from one region with a typical feature set to another expected region. An internal *validation layer* continuously evaluates the success of an knowledge based feature-position prediction and controls the learning process actively by giving reinforcement signals to the *Feature Transition Memory* and the *Location Memory*. So, a reproducable (stable) move from one interesting region to another certain region gains much reinforcement, whereas unstable 'what-where-relations' gain less reinforcement. A more detailed discussion on this topic is given in (Heinke, 1994).

2.5 Behaviour-oriented Object Memory (BOM)

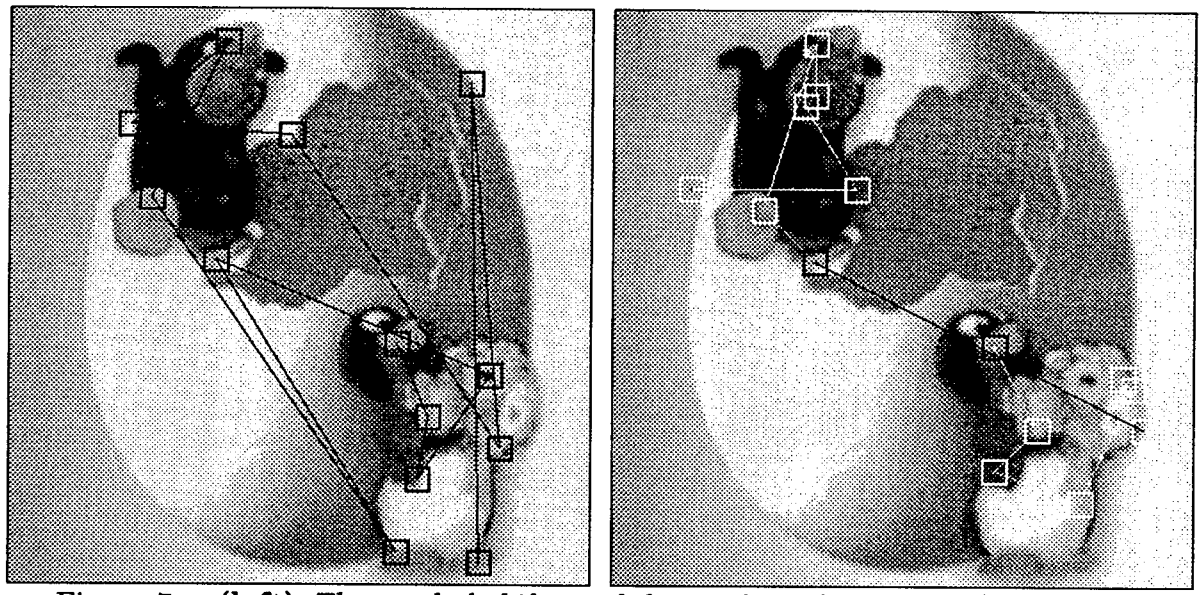

Figure 7: **(left)** *The marked shifting of the window of attention (black frames) shows the course of the data-driven search and the decomposition of the scene into a sequence of striking local input components arranged according to their local conspicuousness. Without any internal knowledge about typical intra-object relations ("what-where-when") our system is not able to establish suitable hypothesis about objects to anticipate an effective scanning. In result, numerous shifts occur between the penguins and the background. The system is not able to select the striking components of the same object successively in time.* **(right)** *This figure illustrates a knowledge controlled scanning process starting from the right penguin. The white frames show foci of attention successfully driven by the self-organized 'what-where-when' object-knowledge acquired during previous analysis of more than 50 scenes (see gallery in Fig. 2). This learning is based on finding out stable intra-object relations within the complex scenes. The remaining black boxes show data-driven moves carried out when the knowledge based 'feature-position' predictions were not successful (results from Heinke, 1994).*

The *Object Memory* is the highest organizational level of our architecture, where both planning and dynamical control of explorative behaviour in visual attention vision take place (see Fig. 5). The memory is activated and driven by the established spatio-temporal sequence of striking input components and can act as a guide in attentional control based on the knowledge already accumulated within the system. The *Object Memory* interacts reciprocally with the *Spatio-temporal Feature Memory* in so-called 'hypothesize-verification-cycles'. In these cycles all activated hypotheses interfere back to the *Spatio-temporal Feature Memory* and try to control the course of the attentional search by generating 'what-where-expectations'. Via this feed-back this memory can search for that input components which would support best one of the activated object-hypotheses. If it is impossible to activate internal hypotheses by a data-driven input sequence, this input has to be accepted by the *Object Memory* as a new sensory situation (e.g. an unknown scene or a new arrangement of known objects within the visual scene) and the sequence or parts of it have to be learned. A concrete approach to a neural architecture suited for such an *Object Memory* is proposed in (Boehme, 1994). In this architecture the components of different temporal scanning se-

quences are mapped into characteristic memory traces within an episodical memory. So, each sequence of decisions on certain striking input components is transferred into a spatio-temporal representation within the *Object Memory*, which can be activated for top-down control in 'hypothesize-verification-cycles'.

3 Simulation Results

For self-organization a behaviour-oriented object-understanding a *gallery of typical scenes* was used (see Fig. 2). In these figures different unknown objects (penguins) are arranged randomly in various locations within the scenes. Also numerous different background situations have been 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 a simple object-understanding. Compared with the original data-driven search dynamics in Figure 7-(left) a drastical modification in the scanning behaviour can be seen in Figure 7-(right). Attentive 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 an acquired *implicit object-understanding* - without a supervised training of all possible objects in a special learning mode before. In this sense, our system was able to demonstrate its ability in changing its *explorative behaviour* in scanning complex scenes. This is the result of transformation reliably *detectable object-specific relations* uncoupled with respect to time at the beginning into more and more *stable temporal relations* by autonomous learning and temporal reshuffling the exploration process.

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] **Boehme, H.-J., Braumann, U.-D., Gross, H.-M. (1994).** A Neural Network Architecture for Episodic Feature Representation. *Proc. of IWK'94, Ilmenau*

[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] **De Valois, R. (1993).** A Multi-Stage Color Model. *Vision Research*, 33, 1053-1065.

[5] **Gross, H.-M., Koerner, E., Boehme, H.-J. (1992).** A Neural Network Hierarchy for Data and Knowledge Controlled Selective Visual Attention. *Proc. ICANN'92*, 825-828

[6] **Heinke, D. & Gross, H.-M. (1993).** A Simple Self-organizing Neural Network Architecture for Selective Visual Attention. *Proc. ICANN'93, Amsterdam*, 63-66, Springer-Verlag.

[7] **Heinke, D. & Gross, H.-M. (1994).** A Neural Network Architecture for Selforganization of Object Understanding. *Proc. IWK'94, Ilmenau*, this volume.

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

[9] **Olshausen, B.; Andersen, C. & van Essen, D. (1993)** A Neural Model of Visual Attention and Invariant Pattern Recognition. *Journal of Neurosciences.* In press.

[10] **Pomierski, T., Gross, H.-M. & Wendt, D. (1993).** A Distributed Multicolumnar System for Primary Cortical Analysis of Real-World Scenes. *Proc. of ICANN'93*, 142-147

[11] **Pomierski, T. & Gross, H.-M. (1994).** Feature Extraction using Self-Organizing Receptive Fields. *Proc. of IWK'94, Ilmenau*, this volume.

[12] **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.