Project no.:	FP7-ICT-217077
Project full title:	Heterogeneous 3-D Perception across Visual Fragments
Project Acronym:	EYESHOTS
Deliverable no:	D3.2
Title of the deliverable:	Object-based top-down selection

Date of Delivery:	03 September 2010	
Organization name of lead contractor for this deliverable:	WWU	
Author(s):	F.H. Hamker, F. Beuth, J. Wiltschut	
Participant(s):	WWU	
Workpackage contributing to the deliverable:	WP3	
Nature:	Other (software module)	
Version:	1.0	
Total number of pages:	22	
Responsible person:	Fred H. Hamker	
Revised by:	Marc Van Hulle	
Start date of project:	1 March 2008	**Duration:** 36 months

Project Co-funded by the European Commission within the Seventh Framework Programme

Dissemination Level		
PU	Public	**X**
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract:

The document explores the question how the brain can bind different visual fragments together to form an object. We are also interested in the question how it can select the appropriate subset of the visual fragments for a searched object in a scene. A visual fragment can encode simple features like edges or more complex parts of an object. All fragments together are perceived as an object. Our approach is that the concept of attention is used to bind these spatially distributed fragments together. The attention process uses feedforward connections to detect the fragments and choose an object. Additional feed-back connections from the object representing cells point back to the fragments representing cells. Their purpose is to reinforce the subset of fragments associated with the object. Solving the binding problem also resolves some object recognition problems like segmentation and localization. Here, we will demonstrate the concept of attention for stereoscopic object recognition in a virtual reality setup, which can be applied to robots in the future. Finally, the report contains the software documentation for the object recognition module.

Contents

1. Executive summary

This document contains the technical report for deliverable D3.2 “Object-based top-down selection using learned bi-directional connections between feature detectors to localize the object of interest in a cluttered 3D scene. Software module.”

The European project “Eyeshots” focuses on the research of a visuo-motor system which is based on the concept of “active and fragmented vision”. It is inspired by the primate brain which actively generates a cognitive interpretation of a perceived scene. It does not encode the scene as pure 2D images or reconstruct real 3D data. Instead, it creates an efficient code in which a scene consists of distributed and loose features, called visual fragments. A visual fragment can represent simple features (like edges or corners) or more complex parts of an object. A subset of fragments is associated with each object and thus forms this object.

We are interested in the question how the brain could select an appropriate subset of the fragments which creates the basis for an object representation. We would like to explore both the feedforward (to bind features to objects) and the feed-back processing (to select features for prior object activations). We assume that the concept of attention is used to bind such distributed fragments together. Our concept of attention uses bi-directional connections. Using the feedforward connections, the object representing cells detect the visual fragments. The feed-back connections project back to reinforce the subset of fragments associated to an object. The connections are learned with a correlation-based learning algorithm, which captures the basics of early human visual perception.

The prior activation of a subset of object selective cells determines which visual fragments are reinforced and thereby the model can select different objects in a cluttered scene depending on the activation pattern of the object selective cells (called the context). Our model of attention addresses the problem of 1) a context dependent selection of objects, 2) localizing an object and 3) parallel recognition and segmentation of an object. Hence we will demonstrate these properties in this deliverable.

The described software module is an object recognition system (ORS). We demonstrate its operation in virtual reality but it is also suitable to be applied to a robot with a stereo head as developed within the Eyeshots consortium. The tasks of the ORS module are to perceive, localize and segment an object in a cluttered scene. The robot's task is represented by some higher level decision processes which determine the relevance of each object in the scene for the current task which is not explicitly modeled. This relevance is projected back by the concept of attention and thus the robot is capable to select the appropriate visual fragments for the task.

2. Introduction

The primate brain actively generates a cognitive interpretation of a perceived scene. It is assumed that it creates an efficient code in which a scene consists of distributed and loose features, called visual fragments. A visual fragment can encode simple features (like edges or corners) or more complex parts of an object. A subset of fragments is associated with each object and thus forms this object. In the following, the terms "visual fragment" and "feature" are used equally. Our approach is that the concept of attention (see 2.1) binds these distributed fragments together. We will first explain this mechanism and how it can be used for solving the problem of binding and selecting features. To be more precise, we would like to investigate the question of how the brain can select an appropriate subset of the features which creates the basis for an object representation. We would like to explore both the feedforward (to bind features to objects) and the feedback processing (to select features for prior object activations).

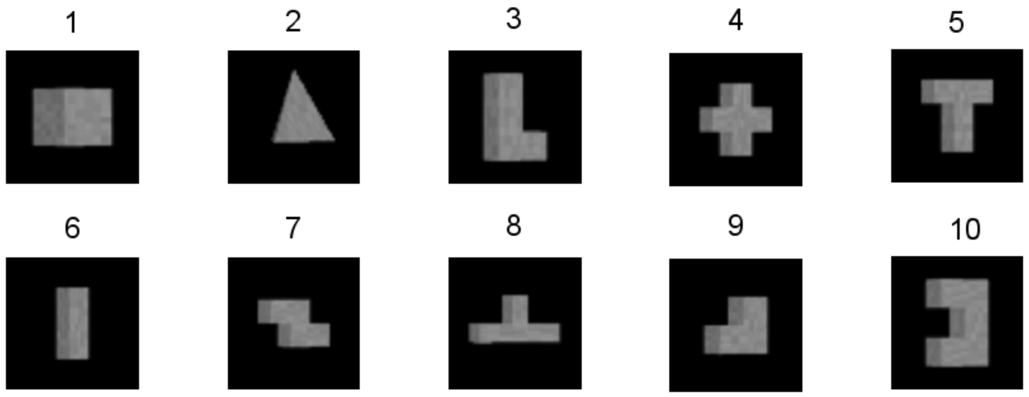


Figure 1: The stimuli consist of 10 different 3D objects.

We address the following three problems:

1. The problem of selecting an object depending on the current context: A task (e.g for a robot) is typically represented by some higher level decision processes which determine the relevance of each object for the current task in the scene. This relevance has to be propagated back to select the correct visual fragments in a scene and filter the distracting ones out. If only a single object is relevant for the current task, we call it the searched object. In this report, we will demonstrate this setup.
2. The problem of localizing an object in a scene: The problem is linked to the first problem and just means to detect the spatial position of the searched object in the scene.
3. The problem of parallel recognition and segmentation an object: For recognizing a searched object in a scene, the object must first be located and segmented, which however is only possible if the object has been recognized as such. Attention can solve this “chicken-and-egg-problem” due to its parallel computation approach.[4]

We will first describe the concept of attention, which is able to achieve object-based top-down selection. In the third chapter, we will present the neuronal network architecture of the object recognition software module and necessary preprocessing steps. We use preprocessing of stereoscopic image data (see the stimuli in Fig. 1) via an edge and depth detection model. We will explicitly focus on explaining the binding of visual fragments to an object (bottom-up) and on object-based top-down selection. The evaluation in the fourth chapter demonstrates how the concept of attention selects different cells dependent of the context and shows the quality of object selectivity. We conclude this document with the software documentation of the object recognition module (ORS) and the equation appendix.

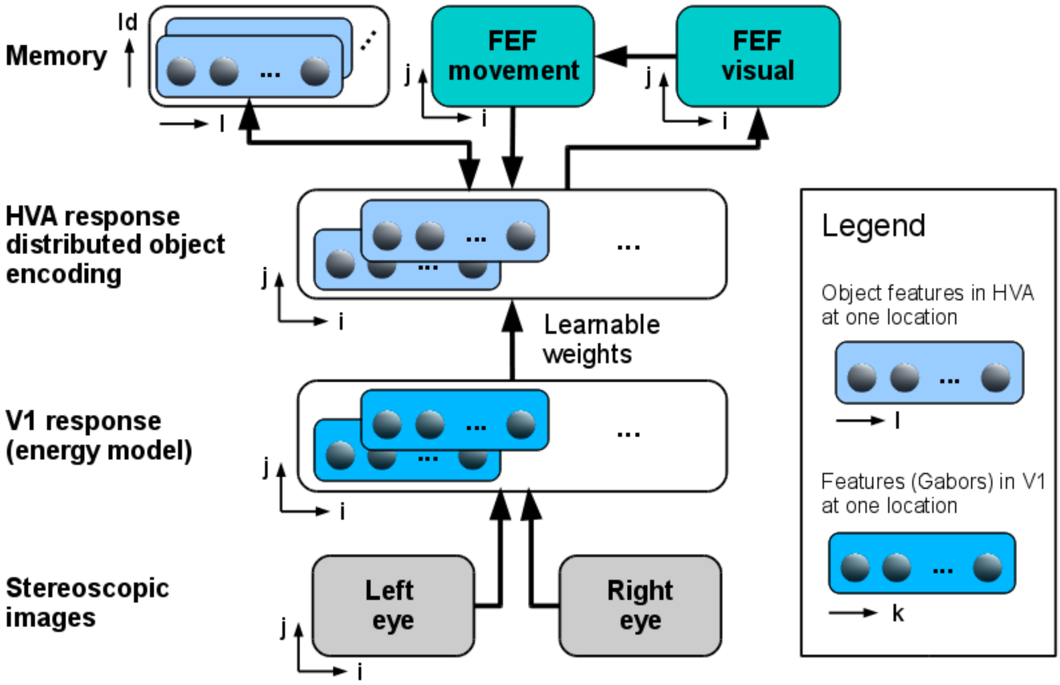

Memory
Id
l
FEF movement
FEF visual
j
i
HVA response distributed object encoding
Learnable weights
V1 response (energy model)
Stereoscopic images
Left eye
Right eye
Legend
Object features in HVA at one location
l
Features (Gabors) in V1 at one location
k

Figure 2: Neuronal network of the stereoscopic object recognition model. The i and j indices correspond to the spatial x and y axis of the images. The index k refers to different Gabor responses and l to different learned features in HVA.

2.1. Concept of attention

Early concepts of visual attention define attention as to focus the processing on a spatially determined part of the image, which has been termed the spotlight of attention. The location of interest is typically determined from conspicuous or salient image features forming the saliency map [7, 9].

Recently, the "spotlight of attention" concept has been expanded to a feature-based approach [4] in which attention emerges from interactions between different brain areas. High level areas hold a template to specify the searched object and this information is propagated backwards to lower level areas. The parallel computation modifies the conspicuity of each descriptor in the system in such a way that the value represents the accumulated evidence. To perceive an object, a combination of several distributed visual features is required. Such binding processes can be well described by concepts of visual attention, illustrated by two continuous subprocesses. The first one operates in parallel over all features and increases the conspicuity of those that are relevant for the searched object, independent of their location in the visual scene. The other subprocess is linked to action plans, e.g. eye movement plans, and combines those fragments, which are consistent with the action plan, typically by their spatial location in the visual scene.

3. Object Recognition System

3.1. Neuronal network architecture - overview

We extend the concept of population-based object representations where we only used simple features [4] by learnable object representations based on local edge detectors. This allows us to detect objects depending on their shape or texture. We measure the recognition performance for all objects by a discrimination value.

In our neuronal model (Fig. 2), we do not consider all the complexity of the visual stream. Rather we simulate an early area (V1) and a high level area (HVA) whose cells can be mapped onto area V2/V4/IT. An object is represented by a distributed code of HVA cells, where a single HVA cell can be interpreted as representing a single view of an object. As input stimuli we use the left and right eye view of 10 different 3D objects (Fig. 1), produced by a raytracer engine [1]. The objects are to some degree similar in their edges and thus subparts of an object can belong to other objects, too. The recognition of these ambiguous subparts results in partially false response of the distributed HVA code. These subparts can also occur in cluttered scenes (the probability increases with the number of objects). Therefore, the difficulty of our problem is comparable to recognize objects in cluttered scenes. The first layer serves as a preprocessing for the object recognition system (ORS), detects stereoscopic edges and disparities via an energy model [10, 12, 13] and is comparable to area V1. This particular energy model [15] uses 56 Gabors with 8 orientations (with a $\frac{\pi}{8}$ step size) and 7 different phase disparity shifts (with $\frac{\pi}{4}$ step size). This area builds a representation of the scene encoding edge information, independent of the right or left view and therefore enables stereo object recognition.

Overlapping receptive fields serve as input for cells of the HVA. We achieve the object selectivity by learning the feedforward weights (V1→ HVA) with a biological motivated learning algorithm and a trace rule (see 3.1.2). The memory generates an attention signal representing the features relevant for the current task, here to search for a specific object in the scene. The Frontal Eye Field (FEF) consists of two areas, the saliency map (called FEF visual / FEFv) and the map encoding the target of the next eye movement (called FEF movement / FEFm). One of the binding processes operates over all locations in HVA and reinforces the features of the searched object (which depends on the current task). The other is achieved by the loop over FEFvisual and FEFmovement. This mechanism reinforces adjacent locations. Both processes use a soft winner takes all competition to decrease the activity of irrelevant features and locations in HVA.

3.1.1. Neuron model

We use a rate coded neuron model which describes the firing rate r of a cell as its average spike frequency. Every cell represents a certain feature (the feature index k refers to are V1, the feature index l points to HVA) at a certain location (with indices i, j). In the following we will omit the location indices for clarity. Consider one location in HVA, each cell in HVA gains excitation (as a weighted sum) from cells of V1 within the receptive field (here a 14x14 patch) and each cell is inhibited by all other HVA cells via Anti-Hebbian inhibition (similar as in [22]).

$$\tau_R \frac{\partial r_l}{\partial t} = \sum_i w_{kl} \cdot r_k^{\text{Input}} - \sum_{l',l' \neq l} f\left(c_{l,l'} \cdot r_{l'}\right) - r_l \quad \text{with: } f(x) = d_{nl} \cdot \log\left(\frac{1+x}{1-x}\right) \tag{1}$$

$f(x)$ gives the non-linear processing. τ_R is the time constant of the differential equation which models the cell's activity. The connection $w_{k,l}$ denotes the strength of the feedforward weight from input (V1) cell k to the output (HVA) cell l. Lateral inhibition is given by the connection weight $c_{l,l'}$ and can differ across the cells due to the Anti-Hebbian learning.

3.1.2. Learning of the object descriptors

Changes in the connection strength between neurons in response to appropriate stimulation are thought to be the physiological basis for learning and memory formation [19]. It has been shown (for the visual system), that learning of correlations within input stimuli is accomplished by a simple principle, the Hebbian law [5]. According to this law connections between neurons (synapses) are strengthen if the corresponding neurons are activated at the same time. Thus, over time cells "learn" to respond to particular inputs. In our model object recognition is achieved by learning the connection weights ($w_{kl}^{\text{V1-HVA}}$) between V1 and HVA. Using a general learning algorithm, that has been shown to capture the features of early visual learning [22], cells from HVA tune themselves to specific features from the set of presented stimuli.

It has been hypothesized that the ventral pathway uses temporal continuity for the development of view-invariant representations of objects [3, 14, 21]. This temporal continuity can be applied using a trace learning rule. The idea is that on the short time scale of stimuli presentation, the visual input is more likely to originate from different views of the same object, rather than from a different object. To combine stimuli that are presented in succession to one another, activation of a pre-synaptic cell is combined with the post-synaptic activation of the previous stimulus using the Hebbian principle. We simulate an appropriate input presentation protocol and the responses of successive stimuli are combined together to achieve a view-invariant representation of an object.

During learning the connection weights $w_{k,l}^{\text{V1-HVA}}$ are changed over time according to:

$$\tau_L \frac{\partial w_{kl}}{\partial t} = \left[r_l^{\text{HVA}} - \tilde{r}^{\text{HVA}}\right]^+ \left((r_k^{\text{V1}} - \tilde{r}^{\text{V1}}) - \alpha_w \left[r_l^{\text{HVA}} - \tilde{r}^{\text{HVA}}\right]^+ w_{kl}\right) \tag{2}$$

$\tilde{r}$ is the mean of the activation over the particular features $\left(e.g., \tilde{r} = \frac{1}{N}\sum_{l=1}^{N} r_l\right)$ and $[x]^+ = \max\{x, 0\}$ and α_w constrains the weights in the same way as the Oja learning rule [11]. The term τ_L is the time constant for learning and thus controls the speed of the learning process. The V1-HVA weights are learned only at a single receptive field (a 14x14 patch of V1) and their values are shared with all other locations in the HVA (weight sharing approach). The learning was performed on small images containing a single stimulus before processing entire scenes (offline-learning).

Lateral connections between cells were learned by Anti-Hebbian learning. The name Anti-Hebbian implies that this strategy is the opposite of the Hebbian learning rule. Similar to the learning of the synaptic connection weights, where the connection between two cells is increased when both fire simultaneously, in the Anti-Hebbian case the inhibition between two cells is strengthened. The more frequent two cells are activated at the same time, the stronger they inhibit each other, increasing the competition among those two cells (l and l'):

$$\tau_C \frac{\partial c_{l,l'}}{\partial t} = r_{l'} \cdot r_l - \alpha_c r_{l'} \cdot c_{l,l'} \tag{3}$$

where τ_C is the learning rate of the Anti-Hebbian weights. Anti-Hebbian learning leads to decorrelated responses and a sparse code of the cell population [2].

3.2. Selection and binding process

The section describes the algorithm which binds fragments to an object (bottom-up) and figures out how fragments are selected depending of the searched object (top-down).

3.2.1. Bottom up processing

We first describe the feature binding process and then the spatial binding process. Through the offline-learning of the object descriptors, the learning algorithm creates a weight matrix from statistical input. This matrix binds the features (here the V1 responses) to an HVA cell. Before the energy responses are processed, we pool these values with a maximum rule over 3x3 elements to decrease the spatial resolution. Every HVA receptive field overlaps highly with its neighboring cells. The pooling rule and the overlapping receptive fields prevent abrupt changes in the response of the cells. At every time step a stimulus is contained by the receptive fields of several neighboring HVA cells. If the stimulus moves from the center to the border of a receptive field, neighboring cells

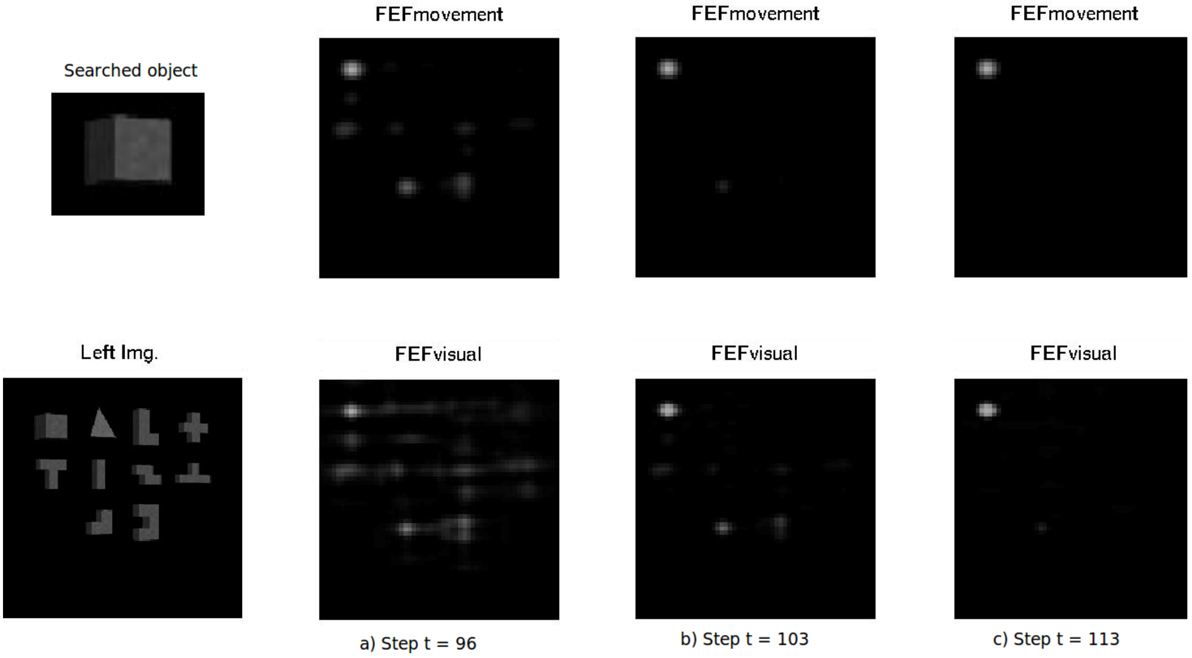


Figure 3: The temporal dynamic of the FEF visual and movement map illustrates the convergence of the system. The current task is to search for the cube in this cluttered scene (both illustrated by the left eye view). The x and y axes correspondent to the spatial x and y axes of the image.

still respond to the stimulus and thus the pooled activation changes smoothly. Hence, the detectors have some tolerance to spatial changes. Low activations are suppressed due to the Anti-Hebbian mechanism (eq. 3). In addition to the spatial resolution, the V1 map has a feature dimension (here over 56 Gabors). Neighboring filters also respond to similar orientations and phase shifts, therefore we also have a smooth response in cases of moving or rotating stimuli. In our special case, the weight matrix binds (spatially and feature based) the V1 responses to a specific HVA cell. As an example, this mechanism shows the concept of binding visual fragment together to encode an object. In general, a stimulus creates specific response patterns in each layer which are bound pairwise between consecutive layers.

The second binding process links spatially close locations together in order to select only a few potential relevant locations in a scene and finally choose the best location as the saccade target. Fig. 3 shows the temporal dynamics of the FEF maps. The task in this example is to search for the cube and pick it out among 10 possible objects.

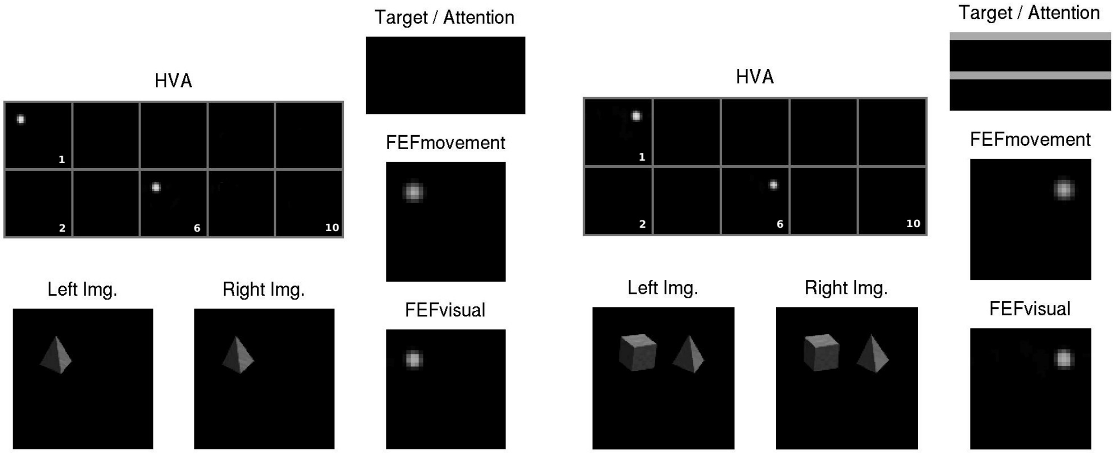
The map of potential locations encodes informations like a saliency map and this area can be mapped to the FEF visual part. The algorithm chooses the maximum over all features for every location in HVA and creates the FEF visual response. Fig. 3(a) demonstrates this situation where the FEF visual part encodes potential locations of the searched object. Then the system reinforces neighboring locations via a Gaussian filter creating the FEF movement response and suppresses other activities by competition. The effect of this reinforcement and competition can be seen in Fig. 3(a) - 3(c) showing that the system filters out unrelevant object candidates over time. After convergence of the system, the FEF movement map contains a Gaussian representing a saccade target (Fig. 3(c)).

Finally, the distributed response of HVA cells are bi-directionally associated to an object identification number (supervised learning). The robot developed within the Eyeshots consortium has not the ability to listen to spoken instruction. Thus, we need a simple mechanism to specify the current task for the robot. We abstract the current job to the task "search for a specific object" and the object is referred by an identification number (id). This mechanism gives us the possibility to search for a specific object via id. However, it is not biologically plausible.

3.2.2. Top-down object selectivity

The top-down selectivity is normally used for a visual search task, i.e. searching for a specific object. The attention signal projects back to the HVA cells modulating their activations. Over all areas of the scene, the attention signal reinforces the firing rates of all cells encoding associated features. All other mechanisms of the feedforward processing work in the same manner: the other feature representing cells in HVA are suppressed due to the Anti-Hebbian mechanism and the loop over FEF visual and FEF movement filters the object spatially.

The robot's task is typically represented by some higher level decision process which determines the relevance of each object in the scene for the current task. This relevance is propagated back and directly filters out irrelevant HVA features. Indirectly this also ignores complete objects. Thus the robot is capable to select the appropriate visual fragments for the current task.



(a)

(b)

Figure 4: The layer activities during the object location experiment. Here the stereoscopic stimuli, the responses of the feature code (with 10 features) in HVA, the attention signal (features on the y-axis) and both FEF areas are shown. Normally, the x and y axes correspondent to the spatial x and y axes of the images. **a)** The system memorizes the target object, the 'tetrahedron', as a combination of HVA cell 1 and 6 and stores the HVA response as an attention signal for b). **b)** The attention signal reinforces the features which represent the 'tetrahedron' and the system detects the target object.

4. Results

We show the ability of context dependent object recognition within a scene containing the searched object and several distractor objects. We also measure the recognition performance for all objects by a discriminating value. The three starting problems are linked together, thus we do not use separate tests for each problem. The solution for the problem "context depending object recognition" also contains a spatial aspect which performs localization and segmentation.

4.1. Object recognition with 2 objects

The task of the robot (the context) determines the relevance of the searched object in the scene. It has to be recognized independent of its position in the image, its rotation or its relative size (for an overview see [16]). Position invariance is achieved in the cortex by pooling over a certain spatial area, which is also considered in our model. The recognition also performs automatically localization and segmentation.

Our first object detection experiment is structured as follows:

1. We present an object alone in a scene without an attention signal (Fig. 4(a)). The model selects the most conspicuous region (the object) and binds the HVA activation to the working memory (which stores the attention signal for each unique object id).
2. We present a black screen to deplete all cell activities in the system.
3. We test the ability to select the target object. We present a cluttered scene (Fig. 4(b)) (here for simplicity with only 10 HVA features and 2 objects). The attention signal encodes the features of the object and reinforces them in HVA. Thereby, the system is able to localize the object (context dependent recognition).

4.2. Object recognition with 10 objects

With the first recognition experiment we have shown that our model can discriminate the searched object from the distractor depending of the current task. In this section, we extend the object recognition to all 10 objects. These objects are to some degree similar in their edges and thus the difficulty of the problem is comparable to cluttered scenes. We test the recognition in two ways, first we visualize the neuronal responses in a location experiment and second we measure the dissimilarities of the attention signals. The second method results in a more precisely measurement as a visual inspection of the neuronal responses. Both tests proof that the object detection works well. In the first examination all objects were detected perfectly and the second measurement reveals good or very good pairwise dissimilarities between all objects.

The first approach uses the same experimental setup like the case of two objects, except that the system contains 50 HVA features. We have arbitrary chosen 50 features and thus the number of features in HVA is 5 times the number of objects. In Fig. 5(a) - 5(d) are shown the neuronal responses for three arbitrary chosen object. Fig. 5(a) demonstrates the memorization of the cube and Fig. 5(b) the location of the cube among the distractors. The object detection works perfectly for each of the 10 objects. The Fig. 5(c) and 5(d) display as examples the recognition for object no. 5 and 7.

4.2.1. Discrimination measurement for 10 objects

To determine the similarity of two feature codes $(\mathbf{r}, \mathbf{s})$ the angle between those two vectors is considered. The lower the value of $d_{TM} \in [0; 1]$ is the more the two vectors show similar cell distributions.

$$d_{TM}(\mathbf{r}, \mathbf{s}) = 1 - \frac{\langle \mathbf{r}, \mathbf{s} \rangle}{|\mathbf{r}|\,|\mathbf{s}|} \quad \text{with: } \dim(\mathbf{r}) = \dim(\mathbf{s}) \tag{4}$$

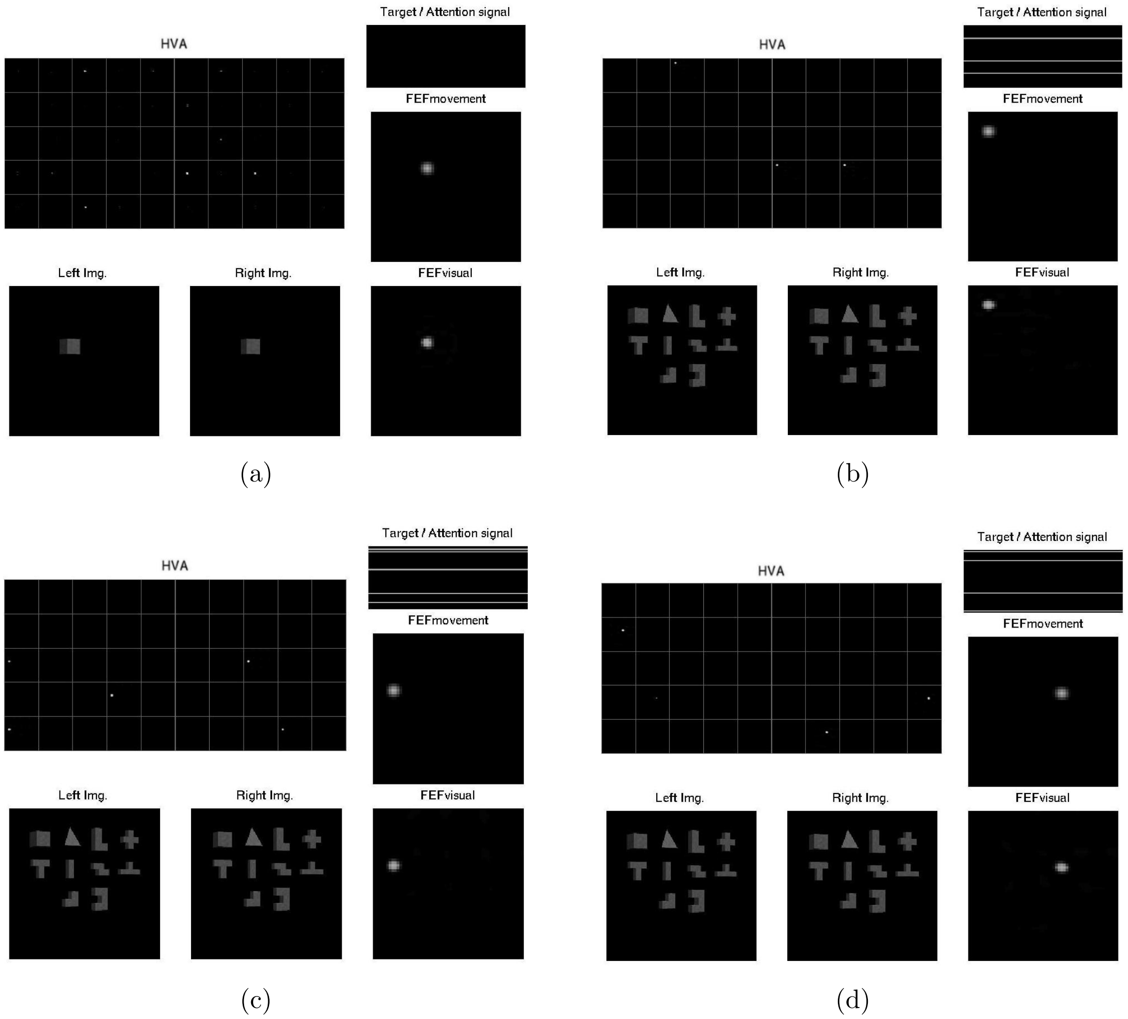
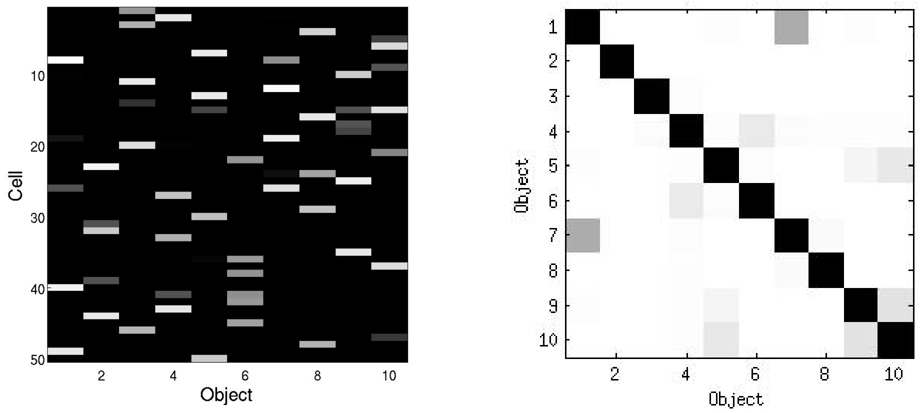


Figure 5: The layer activities during the second object location experiment with 10 object. Here the stereoscopic stimuli, the responses of the feature code (with 50 features) in HVA, the attention signal (features on the y-axis) and both FEF areas are shown. Normally, the x and y axes correspondent to the spatial x and y axes of the images. **a)** The system memorizes the target object, the 'cube' as a combination of three HVA cells. **b)** The attention signal reinforces the cells which encode distributed the 'cube' and the system detects the target object. **c-d)** The location works perfectly for each of the 10 objects which is demonstrated here for two arbitrarily chosen objects.



(a) Object selectivity

(b) Discrimination between objects

Figure 6: **a)** For each object (x-axis) the average firing response (0 dark, 1 bright) of each feature/cell (50 features on the y-axis) is plotted. The average firing response is calculated over all input stimuli that contain the same object. **b)** Using the discrimination value d_{TM}, the similarity of the average response (Fig. 6(a)) to an object is shown here (bright = dissimilar).

Our results show that regardless of the number of different objects and independently of the number of cells (as long as there is at least one cell per object) the model is capable to learn and discriminate all objects. It can be seen that each object is learned by several cells (Fig. 6(a)) and thus an object is characterized by a specific distributed feature code with nearly no overlap to other objects.

An analysis of whether the model is able to discriminate among the objects is shown in Fig. 6(b) using the discrimination value (d_{TM}). Low values (indicated by darker areas) give clue to similar feature codes which would indicate that discrimination between those two objects is impaired. The results show that all objects are very dissimilar in their features and thus are very easy to discriminate. Only object 1 and object 7 show slightly overlapping population codes ($d_{TM} = 0.68$) but the objects can easily be discriminated (see Fig. 6(a)). Although some cells tend to code more than one object the results show that all objects can be discriminated perfectly due to the specific distributed code.

5. Discussion

We have shown how concepts of attention in the human brain could be used to bind loosely distributed visual fragments together and form an object (bottom up). The attention signal also projects back from object selective cells to modulate the cells in HVA (top down). The model uses unsupervised learned feedforward connections between V1 and HVA and supervised learned bi-directional connections between HVA and the memory. Future models should be extended to use additional learned feed-back connections between HVA and V1. Then, the activity of HVA is able to gain-modulate the activity of V1. The spatial resolution of V1 is higher as HVA and thus the model will localize, segment, etc. the object with a higher accuracy.

Another simplification regards the visual stream. We have abstracted the hierarchy of the different areas (V2, V4, IT) to one artificial higher visual area (HVA). We assume that the attention process will work between each pair of two successive areas in the same way as between V1-HVA or HVA-memory. Thus, future models could simulate these processes and they could include more biological aspects. For example a future model could implement realistic receptive field sizes of the areas V2, V4 and IT [18]. In the current system, the field size of an HVA cell is a compromise over a the very large amount of different receptive field sizes.

Our learning algorithm captures the basics of human visual perception, but can be extended to cover complex cell dynamics like calcium traces [17]. We have shown that our system models invariances of the visual cortex. We have focused only on spatial invariance and therefore we will have to extend the model and its learning algorithm to scale and rotation invariance. Most neurons in higher areas have a small rotation and scale invariance, but encode a single view of the object (called view-tuned cells [8]). In further investigations we can compare the properties of the learned cells in the HVA with those view-tuned cells.

A. Appendix: Model details and software documentation

This appendix describes 1) the preprocessing for the object recognition module by the V1 energy model, 2) the software module itself and 3) the equations.

A.1. Model details

After we outline the mathematical notation, we will explain the methods to process the images in order to create the input for the object recognition module.

A.1.1. Mathematical notation

The firing rates of all neurons are labeled with $\boldsymbol{r}$, its elevated term describes the area and its inferior term identifies the neuron indexes (e.g. $\boldsymbol{r}^{\mathrm{HVA}}_{\boldsymbol{i},\boldsymbol{j},\boldsymbol{l}}$). We define the indexes $\boldsymbol{i}$ (X-axis) and $\boldsymbol{j}$ (Y-axis) as spatial ones, index $\boldsymbol{k}$ defines the V1 feature (one of the 56 Gabors) and index $\boldsymbol{l}$ the feature in HVA. All indexes are counted from zero. The firing rates are in range $[0, 1.5]$. A firing rate of 0 represents an inactive neuron. The system have a to avoid firing rates greater than 1 and they should occur very rarely.

A.1.2. Stimuli and energy model

The stereoscopic stimuli are the left and right eye view of 3D objects. We compiled 3D models of cubes to create 10 different objects (see Fig. 1) and chose a raytracer engine [1] developed in WP2 of the Eyeshots project to produce the images.

The eyes are simulated as virtual cameras with a distance of 66mm (like the distance of the human eyes) to each other. They were placed at the positions $(x, y, z) = (-33, 0, 0)$ mm for the left eye and $(+33, 0, 0)$ for the right eye (all positions are described as (x, y, z) vectors and all measurements are in millimeter). The filter bank [15] implements an energy model (see [10, 12, 13]) using 56 Gabors with 8 orientations (with a $\frac{\pi}{8}$ step size) and 7 different phase disparity shifts ($\pi \cdot \{-0.75, -0.5, \ldots, 0.75\}$). The envelope size of each Gabor filter is 11x11 pixel. The system can detect disparities in a range of about $[-1.5, +1.5]$ pixel and thus the 3D world arrangement has to meet these constraints. The receptive field of the object selective cells should have at least a size comparable to a complex cell in V2 (see [6]) and therefore we had to select an appropriate receptive field size of HVA. We have chosen an aperture angle of 3° which encloses a whole object resp. 4.4° for an object with background (resulting in a stimulus shown in Fig. 1, rendered at 52×52 pixel).

A.1.3. Size of the neuronal areas

The sizes of each map depends mainly on the size of the stereoscopic images and the number of objects. Here we show the sizes for images of 220 × 220 pixel and 10 objects.

Name	Width/X	Height/Y	Features
Image	220	220	1
V1	210	210	56
V1 pooled	70	70	56
HVA RF	14	14	-
HVA	57	57	50
Attention	1	1	50
FEFvis	50	50	1
FEFmov	50	50	1

The term HVA RF specifies the spatial receptive field size of a HVA cell with an overlap of 13 pixel. The number of features in HVA should be around 5 or 10 times the number of objects. We have arbitrary chosen 50 features.

A.2. Software documentation

This section describes the algorithm and the interfaces of the object recognition software module (called ORS). In the context of the project "Vergence-Version Control with Attention Effects" (VVCA) we developed a Matlab/Simulink implementation operating in a virtual reality. To apply the software to the robot of the Eyeshots project, we are developing a C++ implementation which uses the neuronal network framework ANNarchy [20, Appendix software module]. The C++ and the Matlab/Simulink implementations have identical interfaces. We here refer to the Matlab/Simulink implementation of the algorithm. The complete package is available to the project partners upon request. Please keep in mind that this software package only contains the object recognition module and therefore only works on the top of the V1 filter bank which is not part of this module.

A.2.1. Software interfaces

Input into the ORS:

1. *V1Out*: A 2D Matrix of the size: (imgX·ImgY) × number of V1 features.
2. *ObjectNo*: Holds the identification number (id) of the searched object, which depends on the objects in the world and the loaded memory file. In the most cases the reference numbers are: 1=cube, 2=tetrahedron, ...
3. *OrsMode*: When equals to '0', the ORS searches for the most conspicuous object in the scene and memorizes its shape as the object number *ObjectNo*. When equals to

'1', the ORS uses the memorized features (shape) from *OrsMode=0* and searches the object again.

4. *StartTrigger*: The object recognition starts every time the trigger changes (0→1 or 1→0). For a single ORS run put a constant 1 here.

Output of the ORS:

1. *FEFm*: The FEF movement map. It encodes the saccade target position usually in form of a 2D Gaussian function.
2. *EndTrigger*: If the ORS has done its task (the HVA activations are stable), the trigger is toggled (0→1 or 1→0).

A.2.2. Algorithm

The main functionality is provided through algorithm 1. It is called in every time step of the simulation. The numbers in round brackets refer to the equations of chapter A.3.

The software stores the features for each object in memory where the object is referred by its id. The memory is saved after the end of the experiment and is automatically loaded at the next start. With this mechanism the user can directly search for objects without presenting the objects first.

Algorithm 1: For each time step

```
if StartTrigger toggle then
    v1pool := pooling (V1out)        (eq. 5)
    HVAff := calculate the feedforward response of HVA from v1pool      (eq. 6)
    if OrsMode == 0 then
        attention := 0
    else
        attention := memory[ObjectNo]
orsStep()
if $|\max\{r_t^{HVA}\} - \max\{r_{t-1}^{HVA}\}| < 0.02$ then
    toggle EndTrigger
    if OrsMode == 0 then
        // Memorize object
        {p,q} := x,y position of maximal FEF movement activity        (eq. 16)
        bestHVA := [0, 1]-normalized feature vector of $r^{\mathrm{HVA}}$ at position {p,q}
        (eq. 17, 18)
        set all values < 0.5 of bestHVA to 0        (eq. 19)
        store bestHVA as memory[ObjectNo]
```

Function `orsStep`

```
// Calculate HVA response
$exc$ := calculate excitation        (eq. 7)
$inh$ := calculate lateral inhibition        (eq. 8)
// Use euler method to calculate the differential equation of HVA
$r^{\mathrm{HVA}} = r^{\mathrm{HVA}} + 1/\tau_R \cdot (-r^{\mathrm{HVA}} + exc - inh)$        (eq. 8)
$r^{\mathrm{HVA}}$ = constraining $(r^{\mathrm{HVA}})$        (eq. 10, 11)
// Calculate FEF visual and movement responses
$r^{\mathrm{FEFv}}$ := maximum over the features at each position of $r^{\mathrm{HVA}}$        (eq. 12)
$r^{FEFm}$ := convolve $r^{\mathrm{FEFv}}$ with a Gaussian and apply a soft-wta rule        (eq. 13)
```

A.3. Equations

The term r describes a firing rate and the terms i, j, k, l describe the map indexes. The section A.1.1 explains the usage. A standard value is specified for each parameter. The values were manually chosen among the best values of the simulations.

A.3.1. Higher visual area (HVA)

$$\begin{aligned}
\forall i,j \;:\quad r^{\mathrm{V1}}_{i,j,k} &= \max_{i'=\{3i..3i+3\},\; j'=\{3j..3i+3\}} \left\{ r^{\mathrm{V1}}_{i',j',k} \right\} && (5)\\
b^{\mathrm{HVA}}_{i,j,l} &= \sum_{i'=i}^{i+z-1} \sum_{j'=j}^{j+z-1} w^{\text{V1-HVA}}_{i',j',k,l} \cdot r^{\mathrm{V1}}_{i',j',k} && (6)\\
g\left(b^{\mathrm{HVA}}, a, r^{\mathrm{FEFm}}\right)_{i,j,l} &= b^{\mathrm{HVA}}_{i,j,l} \cdot (1 - \max\{a\} + a_l) \cdot \\
&\quad \left(1 - \max\{r^{\mathrm{FEFm}}\} + r^{\mathrm{FEFm}}_{i,j}\right) && (7)\\
\tau_R \frac{\partial r^{\mathrm{HVA}}_{i,j,l}}{\partial t} &= g(b^{\mathrm{HVA}}, a, r^{\mathrm{FEFm}})_{i,j,l} - \sum_{l',l'\neq l} f\left(c^{\mathrm{HVA}}_{l,l'} \cdot r^{\mathrm{HVA}}_{i,j,l'}\right) - r^{\mathrm{HVA}}_{i,j,l} && (8)\\
f(x) &= d_{ln} \cdot \log\left(\frac{1+x}{1-x}\right) && (9)\\
\forall\; r^{\mathrm{HVA}}_{i,j,l} < 0 \quad &: \quad r^{\mathrm{HVA}}_{i,j,l} = 0 && (10)\\
\forall\; r^{\mathrm{HVA}}_{i,j,l} > 1 \quad &: \quad r^{\mathrm{HVA}}_{i,j,l} = 0.5 + \frac{1}{1 + e^{(-3.5(r^{\mathrm{HVA}}_{i,j,l} - 1))}} && (11)
\end{aligned}$$

with neuron firing rate r^{HVA} of area HVA, feedforward influence b^{HVA}, feedforward weights $w^{\text{V1-HVA}}$ between V1 and HVA, attention a, attention modulated excitation g, receptive field size $z = 14$, lateral inhibition in area HVA c^{HVA}, time constant $\tau_R = 10$ and non linearity function $f(x)$ with $d_{ln} = 0.8$.

A.3.2. Frontal eye field (FEF)

$$\begin{aligned} r_{ij}^{\mathrm{FEFv}} &= \max_{l} \left\{ r_{i,j,l}^{\mathrm{HVA}} \right\} && (12) \\ r_{ij}^{\mathrm{FEFm}} &= h(r^{\mathrm{FEFv}} * G) && (13) \\ h(X) &= \left(\frac{X^p}{\max\{X^p\}}(1+c) - c \right) \cdot \max\{r^{\mathrm{FEFv}}\} && (14) \\ \text{with:}\quad x &= [-K.. + K] \\ y &= [-K.. + K] \\ G &= e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)}, \quad \text{with: } \sigma = 0.25 \cdot K && (15) \end{aligned}$$

The term $r^{\mathrm{FEFv}} * G$ describes the convolution from r^{FEFv} and G. The kernel size of the Gaussian is K. The function h is a soft WTA (Winner Takes All) function which increases the signal contrast (power rule factor $p = 1.8$). Additionally the function preserves the maximum value and decreases the minimum by a global inhibition mechanism, controlled by the term $c = 0.1 \cdot \max\left\{r^{\mathrm{FEFm}}\right\}$.

A.3.3. Memorization

$$\begin{aligned} \{p, q\} &= \underset{i,j}{\operatorname{argmax}} \left\{ r_{i,j}^{\mathrm{FEFm}} \right\} && (16) \\ \forall\, l: \quad a_l &= r_{p,q,l}^{\mathrm{HVA}} && (17) \\ a &= \frac{a - \max\{a\}}{\max\{a\} - \min\{a\}} && (18) \\ \forall\, l, a_l < 0.5 \quad &: \quad a_l = 0 && (19) \\ & && (20) \end{aligned}$$

The coordinates p and q specify the point of the maximum FEF movement responses which encodes the position of the searched object.

References

[1] N. Chumerin. Nikolay Chumerin's myRaytracer, 2009.

[2] P. Földiàk. Forming sparse representations by local anti-hebbian learning. *Biol Cybern*, 64:165–170, 1990.

[3] P. Földiàk. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[4] F. H. Hamker. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Comp Vis Image Understand*, 100:64–106, 2005.

[5] D. O. Hebb. *Organization of Behavior*. John Wiley and Sons, 1949.

[6] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol*, 28:229–89, March 1965.

[7] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res*, 40(10-12):1489–506, 2000.

[8] N. K. Logothetis, J. Pauls, and T. Poggio. Spatial reference frames for object recognition tuning for rotations in depth. In *AI Memo 1533, Massachusetts Institute of Technology*, pages 12–0. MIT Press, 1995.

[9] R. Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, University of Geneva, 1993.

[10] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, 249(4972):1037–41, August 1990.

[11] E. Oja. A simplified neuron model as a principal component analyzer. *J Math Biol*, 15(3):267–273, 1982.

[12] N. Qian. Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6(3):390–404, 1994.

[13] J. C. A. Read and B. G. Cumming. Sensors for impossible stimuli may solve the stereo correspondence problem. *Nat Neurosci*, 10(10):1322–8, October 2007.

[14] E. T. Rolls and S. M. Stringer. Invariant object recognition in the visual system with error correction and temporal difference learning. *Network*, 12(2):111–129, May 2001.

[15] S.P. Sabatini, G. Gastaldi, F. Solari, J. Diaz, E. Ros, K. Pauwels, M.M. Van Hulle, N. Pugeault, and N. Krüger. Compact and accurate early vision processing in the harmonic space. In *International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona*, 2007.

[16] T. Serre. *Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines*. PhD thesis, Massachusetts Institute of Technology, 2006.

[17] Harel Z Shouval, Gastone C Castellani, Brian S Blais, Luk C Yeung, and Leon N Cooper. Converging evidence for a simplified biophysical model of synaptic plasticity. *Biol Cybern*, 87(5-6):383–91, December 2002.

[18] A. T. Smith, K. D. Singh, A. L. Williams, and M. W. Greenlee. Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cerebral Cortex*, 11(12):1182, 2001.

[19] L. R. Squire and E. R. Kandel. *Memory: From Mind to Molecules.* Roberts & Co Publ, 2008.

[20] J. Vitay and F. H. Hamker. Eyeshots Deliverable D3.3a - Working Memory Model. 2010.

[21] G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Prog Neurobiol*, 51(2):167–194, Feb 1997.

[22] J. Wiltschut and F. H. Hamker. Efficient coding correlates with spatial frequency tuning in a model of v1 receptive field organization. *Vis Neurosci*, 26:21–34, 2009.