Supplementary material to the article “Attention as cognitive, holistic control of the visual system”

Frederik Beuth, Fred H. Hamker

Technische Universität Chemnitz, Artificial Intelligence,
Strasse der Nationen 62, 09111 Chemnitz, Germany
{beuth,fhamker}@cs.tu-chemnitz.de
http://www.tu-chemnitz.de/informatik/KI/

Abstract This supplementary material comprises the full mathematical description of the proposed visual attention model, and illustrates the learning of the object representations. The document is an excerpt from the upcoming doctoral thesis of the first author, Frederik Beuth.

1 Mathematical description of the model

After we outline the mathematical notation, we will illustrate the equations for each area.

Mathematical notation The firing rates of all neurons are labeled with $\boldsymbol{r}$, whereby an elevated term describes the area and an inferior term identifies the neuron indices (e.g. $\boldsymbol{r}^{\mathrm{V1}}_{\boldsymbol{d,i,x}}$). We define the index $\boldsymbol{x}$ as spatial one which contains the Y ($\boldsymbol{x_1}$) and X ($\boldsymbol{x_2}$) - coordinates of an location. The index $\boldsymbol{d}$ defines the channel, which can be red-green (RG), blue-yellow (BY), or orientation (O). The third kind of index is $\boldsymbol{i}$, which define the $\boldsymbol{i}$th feature in the population at a certain position in a certain channel. Indices with the symbol ’ (e.g. $\boldsymbol{i'}$) indicate local loop indices which are used for example by maximum or sum operations. All indices are counted from one.

Connections are modeled via two variables, a weight matrix w (which is normalized to 1) controlling the connectivity and a scalar v controlling separately the amplitude of the integrated signal. Weight matrices connecting *area1* to *area2* are termed as $w^{\text{area1-area2}}_{x,x'}$ with the current post-synaptic neuron x and the pre-synaptic neuron x'. Weight matrices for a suppressive connection are termed according to their function, e.g. w^{SUR} for surround suppression. The scalar v is indexed similarly.

Mathematical definitions

- The term $\#x$ returns the number of elements of an area.
- The function $f_1(x)$ defines a half-rectification of x:
  $f_1(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$

- The function $f_2(x)$ constrains x to a range between 0 and 1:
  $$f_2(x) = \begin{cases} 1 & x \geq 1 \\ x & 0 \leq x \leq 1 \\ 0 & x < 0 \end{cases}$$
- The function $x' \in \mathrm{RF}(x, area)$ returns all pre-synaptic neuron indices x' in $area$ that are in the receptive field of the current post-synaptic neuron x.
- The response of area FEFvm is always used as a mean among its features. Thus we define: $\bar{r}_x^{\mathrm{FEFvm}} = \frac{1}{\#i} \sum\limits_{i'} r_{i',x}^{\mathrm{FEFvm}}$
- The function $\mathfrak{g}$ represents a two-dimensional Gaussian function, whereby a denotes the amplitude, x' the center, and σ the standard deviation. The envelope is typically chosen as $3\sigma_1 \times 3\sigma_2$ as a compromise between calculation speed and sampling precision. The Gaussian is typically centered, thus x' is chosen as half of the envelope.

$$\mathfrak{g}(x, a, \sigma) = a \cdot \exp\left(-\left(\frac{(x_1 - x_1')^2}{2\sigma_1^2} + \frac{(x_2 - x_2')^2}{2\sigma_2^2}\right)\right)$$

Early visual processing - retina Visual processing starts with the absorption of light in the retina by cones and rods. We only consider daylight vision in our model, therefore we simulate no rods, but the three cone types L, M, S, corresponding to long(L), middle(M) and short(S) wavelength. Their peak absorption wavelengths (λ) are defined at $\lambda_L = 560$, $\lambda_M = 530$ and $\lambda_S = 420$ with relative strength v of $v_L = 70\%$, $v_M = 86\%$, $v_S = 100\%$ (Bowmaker and Dartnall, 1980; Stockman and Sharpe, 2000). The human color perception can be approximated by a particular color space model, the LMS color space. We evaluated several approaches modeling the LMS color space and found that the newest international LMS standard, CAT02 (Moroney et al., 2002) in combination with a gamma correction of RGB images, represents very well the absorption properties of the cones in the human eye. We use the MATLAB implementation of Getreuer (2010) which initially transforms the RGB input images to an intermediate XYZ color space and corrects the gamma, and then transforms the result to the LMS color space.

Early visual processing - LGN The most common types of LGN cells (Wiesel and Hubel, 1966) are simulated by our model: L-M color-opponency cells in the parvocellular layers of LGN, S-(L+M) color-opponency cells in the koniocellular layers, and L+M luminance-opponency cells in the magnocellular layers. The terms L, M, S refer to the cone responses of the retina (Gegenfurtner, 2003). We model only cell types which are functionally relevant and observed in the majority of physiological data sets (Chatterjee and Callaway, 2003; Dacey, 2000; Sincich and Horton, 2005; Wiesel and Hubel, 1966). An overview of all cell types and their distribution can be found in Wiesel and Hubel (1966) and Chatterjee and Callaway (2003).

The first cell type, the L-M color-opponency cell, has a center-surround receptive field structure whereby center and surround are driven by different cone types. The cell type is also called single-opponent, midget, or type I cell (according to Wiesel and Hubel (1966)). Its receptive field is modeled by a difference of Gaussians (DoG, Eq. 1), whereby σ_{c} denotes the standard deviation of the center Gaussian and σ_{s} of the surround Gaussian. The center signal is convolved with the positive part of the DoG and the surround signal with the negative part of the DoG (Eq. 4).

$$DoG(\sigma_{\mathrm{c}}, \sigma_{\mathrm{s}})_{x'} = \mathfrak{g}(x', 1, \sigma_{\mathrm{c}}) - \mathfrak{g}(x', 1, \sigma_{\mathrm{s}})\,, \quad \sigma_{\mathrm{c}} < \sigma_{\mathrm{s}} \tag{1}$$

$$DoG_c = a_1\ (DoG)^+ \tag{2}$$

$$DoG_s = a_2\ (-DoG)^+ \tag{3}$$

$$r = DoG_c * r_{\mathrm{c}} - DoG_s * r_{\mathrm{s}} \tag{4}$$

Whereby the factors a_1, a_2 normalize DoG_c and DoG_s to a sum of 1, and the symbol $*$ denotes convolution.

Four different subtypes of L-M color-opponency cells were modeled (Eq. 5 - 8), two ON and two OFF cell types. ON cells have an excitatory center driven by L cones (or M cones respectively), and an inhibitory surround driven by M (or L) cones: L^+M^- and M^+L^-. Analogously, OFF cells have an inhibitory center and an excitatory surround: L^-M^+, and M^-L^+.

$$r^{\text{L+M-}} = r^{\text{L}} * DoG_c - r^{\text{M}} * DoG_s \tag{5}$$

$$r^{\text{M+L-}} = r^{\text{M}} * DoG_c - r^{\text{L}} * DoG_s \tag{6}$$

$$r^{\text{L-M+}} = -r^{\text{L}} * DoG_c + r^{\text{M}} * DoG_s \tag{7}$$

$$r^{\text{M-L+}} = -r^{\text{M}} * DoG_c + r^{\text{L}} * DoG_s \tag{8}$$

The size of the surround is chosen as 0.45°, according to the receptive field size data provided by Smith et al. (2001, Fig 9). The center of type I cells is roughly 4 times smaller as the total field Wiesel and Hubel (1966), thus we choose 0.11°. We model each region by a 2D-Gaussian with an extent of 3σ, therefore, we choose a standard deviation of 0.15° for the surround and 0.0375° for the center. As the calculations are executed in image dimensions, we have to convert these values to pixels. We choose to map 40 pixels to 1°, which results in $\sigma_{\text{parvo-c}} = 1.5$ and $\sigma_{\text{parvo-s}} = 6$ pixels. Therefore, the size of the surround envelope is 19 pixels, corresponding to $3\,\sigma_{\text{parvo-s}}$ and a rounding to the next odd number. The total receptive field is per definition equal to the surround region, thus it has also an envelope size of 19 pixel.

The second cell type, the S-(L+M) color-opponency cell, reacts to differences between the S cones and the combined L & M cones, hence it reacts roughly to blue/yellow contrasts. The type is also called bistratified or type II cell (Wiesel and Hubel, 1966). Both cones drives the same spatial part of the receptive field, hence there exist no center-surround separation in the field. We model both parts with a Gaussian (Eq. 9 - 12). The receptive field size is similar to the size of the L-M cells (Wiesel and Hubel, 1966), thus we choose identically $\sigma_{\mathrm{S/LM}} = 6$.

$$r^{\text{LM}} = (r^{\text{L}} + r^{\text{M}})/2 \tag{9}$$

$$G_{\mathrm{S/LM}} = \mathfrak{g}(x', a, \sigma_{\mathrm{S/LM}}) \tag{10}$$

$$r^{\text{S/LM}} = r^{\text{S}} * G_{\mathrm{S/LM}} - r^{\text{LM}} * G_{\mathrm{S/LM}} \tag{11}$$

$$r^{\text{LM/S}} = r^{\text{LM}} * G_{\mathrm{S/LM}} - r^{\text{S}} * G_{\mathrm{S/LM}} \tag{12}$$

Whereby the factor a normalizes $G_{\mathrm{S/LM}}$ to a sum of 1.

The third cell type, the L+M luminance-opponency cell, reacts to luminance contrasts and has a center-surround separation in their receptive fields. The type is also called parasol or type III cell (Wiesel and Hubel, 1966). It is located in the magnocellular pathway and it is assumed that their information is later combined in V1 to drive edge detecting cells. For simplicity, we do not model L+M cells, instead we simulate directly V1 cells detecting luminance edges (next section). We employ the standard approach, Gabor filters on a grayscale image, to detect such edges. The grayscale image is created from the RGB values via the MATLAB function *rgb2gray* (Eq. 13).

$$r^{\text{GRAY}} = 0.2989\, r^{\text{R}} + 0.5870\, r^{\text{G}} + 0.1140\, r^{\text{B}} \tag{13}$$

Early visual processing - primary visual cortex V1 Our model simulates color and form encoding V1 simple cells which are grouped into three channels: a red-green (L-M), a blue-yellow (S-LM) and an orientation (O) channel. Each channel contains 8 feature cells, which represent different grades of the color opponency in the L-M and S-LM channels, and different oriented edges in the O channel.

For the L-M channel, the cells span a feature space (Hamker, 2005a) between L-active cells ($L^{+}M^{-}$, $M^{-}L^{+}$; Eq. 15) and M-active cells ($M^{+}L^{-}$, $L^{-}M^{+}$; Eq. 16). Each cell prefers a certain feature, namely a particular activity of L or M cells. This preference is modeled by Gaussian tuning functions (Eq. 14) with standard deviation $\sigma = 0.092$ and a mean μ. The channel contains at first four cells preferring L values: $\mu_{i=[1,4]} = \{1.0, 0.77, 0.54, 0.31\}$, and then four preferring M values: $\mu_{i=[5,8]} = \{0.31, 0.54, 0.77, 1.0\}$.

The cells of the blue-yellow channel are modeled similar (Eq. 17 and 18).

$$H(v, i) = \exp\left(-\frac{(f_1\,(v) - \mu_i)^2}{2\sigma^2}\right) \tag{14}$$

$$r^{\text{V1S}}_{d=1,\, i=[1,4],\, x} = H(v^{\text{Lgn}} \max\{r_x^{\text{L+M-}},\, r_x^{\text{M-L+}}\}, i) \tag{15}$$

$$r^{\text{V1S}}_{d=1,\, i=[5,8],\, x} = H(v^{\text{Lgn}} \max\{r_x^{\text{M+L-}},\, r_x^{\text{L-M+}}\}, i) \tag{16}$$

$$r^{\text{V1S}}_{d=2,\, i=[1,4],\, x} = H(v^{\text{Lgn}}\, r_x^{\text{S/LM}}, i) \tag{17}$$

$$r^{\text{V1S}}_{d=2,\, i=[5,8],\, x} = H(v^{\text{Lgn}}\, r_x^{\text{LM/S}}, i) \tag{18}$$

Whereby $v^{\text{Lgn}} = 3$ denotes a scaling factor.

In the O channel, the cells filters the luminance signal r^{GRAY} to detect 8 different oriented edges in 45 degree spacing (Eq. 19). The receptive field of such an oriented-edge cells is modeled by a discretized 2D-Gabor filter (Eq. 20- 22) as proposed by Jones and Palmer (1987).

$$\theta_i = \{0\pi,\ 0.25\pi,\ ...\ 1.75\pi\} \tag{19}$$

$$X_{1,i} = x_1 \cos\theta_i + x_2 \sin\theta_i \tag{20}$$

$$X_{2,i} = -x_1 \sin\theta_i + x_2 \cos\theta_i \tag{21}$$

$$G_{i,x} = A \cdot exp\left(-\left(\frac{X_{1,i}^2}{2\sigma_1^2} + \frac{X_{2,i}^2}{2\sigma_2^2}\right)\right) \cdot \cos(2\pi f X_{1,i} + \psi) \tag{22}$$

$$r^{\text{V1S}}_{d=3,\,i,\,x} = r^{\text{GRAY}} * G_i \tag{23}$$

Whereby $\theta = [0, 2\,\pi)$ represents the orientations, $f = \frac{1}{18}$ the spatial frequency, $\psi = \frac{\pi}{2}$ the phase offset, and $\sigma_1 = 4.5, \sigma_2 = 18$ the standard deviation. Each Gabor is individually normalized by a factor A to ensure that the sum of the positive part of the Gabor is 1. We choose an envelope size of 19 pixels, identical to the size of the LGN cells, thus the discretization points x_1, x_2 run from -9 to $+9$.

The V1 simple cell responses are spatially pooled to V1 complex cell responses. This increases the spatial invariance and decreases the resolution of V1. A complex cell response results from pooling over an area of 10 × 10 simple cells with identical features (Eq. 25) In addition, response differences are enhanced via a non-linearity (Eq. 26). Both approaches are similar as in Antonelli et al. (2014), except that we implement the pooling operation more sophisticatedly by a weighted sum instead a non-weighted maximum. The non-weighted maximum operation leads to discretization errors at the borders of the pooling area. We solve this by a weighted sum with a kernel containing strong weights inside the pooling area, but also weak weights outside it. The latter smooths out the response at the borders and avoids the problem. As kernel, we use a Lanczos3 kernel (Eq. 24) as it meets the requirements and is an often-used standard approach for decreasing resolutions (Turkowski and Gabriel, 1990).

$$K_x = \begin{cases} 1 & x' = 0 \\ (a \sin(\pi\, x')\ \sin(\pi\, x'/a))/(\pi^2\, x'^2) & 0 < |x'| < a \\ 0 & x' \geq a \end{cases}, \tag{24}$$

$$\text{with}\ :\ a = 3, \quad x' = x/(2 \cdot 10)$$

$$R_{d,i,x} = r^{\text{V1S}}_{d,i} * K * K^T \tag{25}$$

$$r^{\text{V1C}}_{d,i,x} = R^{p_{V1C}} \tag{26}$$

Where $p_{V1C} = 2.5$ parameterized the non-linearity, and $*$ denotes convolution, executed separately for each channel d and feature i.

Higher visual area (HVA) - layer 4 The higher visual area (HVA) represents, as an abstract entity, a high-level visual area like the fourth visual cortex (V4)

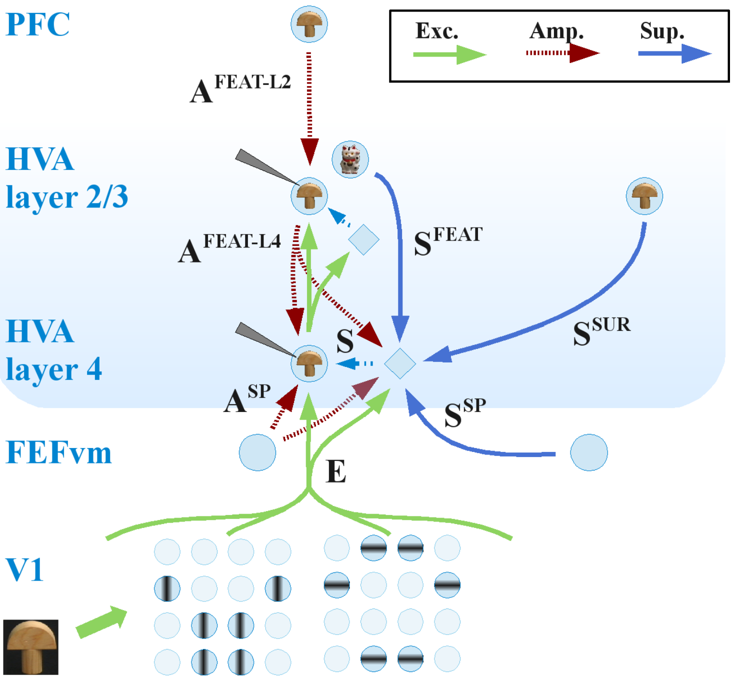

Figure 1. Detailed illustration of the higher visual area (HVA), when recognizing an object. The stimulus excites a specific spatial response pattern in each feature plane in V1, exemplary shown by planes encoding vertical and horizontal edges. Excitation (E) of HVA is calculated from this pattern via a weighted sum. The remaining of the figure shows the connectivity and influences on a single cell in HVA layer 4 and in layer 2/3, indicated by the electrode symbols. The layer 4 cell receives feedforward excitation from V1 (E), feature-based amplification ($A^{FEAT-L4}$) from layer 2/3, spatial amplification (A^{SP}) from FEFvm, and suppression from an associated interneuron (S). The interneuron receives several sources of suppression: the feedforward excitation of its associated neuron (E), dissimilar objects in layer 2/3 (S^{FEAT}), similar objects in the surround in layer 2/3 (S^{SUR}), and other retinotopic locations in the FEFvm (S^{SP}). The layer 2/3 cell receives excitation from layer 4, suppression from its associated interneuron (not shown), and amplification from PFC ($A^{FEAT-L2}$).

or the inferior temporal cortex (IT). V4 represents complex shapes or parts of an object (Cadieu et al., 2007; Hegdé and Van Essen, 2007; Pasupathy and Connor, 2002). Cells of IT react to whole objects (Kriegeskorte, 2009; Op de Beeck et al., 2001; Serre et al., 2007; Tanaka, 1996) or views of whole objects (Logothetis et al., 1995). In the object localization task, HVA contains such view-tuned cells.

HVA is implemented by the mechanistic microcircuit model of attention (Fig. 1). We focus in this section on the embedding of the microcircuit in the larger system-level model, and would like to refer the reader to Beuth and Hamker (2015) for its anatomical and neurophysiological background. The microcircuit model is differently parametrized than in Beuth and Hamker (2015) to modify it for the system-level model and to adapt it to the object localization task. As first primary change, we strengthen the feature-based amplification from PFC to HVA ($A^{FEAT-L2}$) from $v^{\text{PFC-HVA2}} = 0.5$ to 1.5 as the value results in a much higher localization performance, especially on noisy and real-world scenes. To balance out amplification and suppression, we also strengthen the feature-based suppression (S^{FEAT}) by increasing the non-linearity in the suppression from x^2 to x^5. As second major change, we add spatial suppressive connection (S^{SP}) from the frontal eye field. The signal is necessary to suppress the location of distractors, and is a required part of the target selection process within the recurrent loop between FEF and HVA. Thus, the modification results from combining the novel FEF model with the microcircuit model in HVA.

The model responses are simulated by the following equations. The firing rates of HVA and FEF are simulated via ordinary differential equations using the Euler method (Atkinson, 1989) with time step $h = 1ms$, and are constrained to $[0, 1]$ by the function $f_1(x)$. Cell indices d, i, and x are omitted if all terms within an equation refer to the same cell.

$$\tau^{\text{HVA4}} \frac{\partial r^{\text{HVA4}}_{d,i,x}}{\partial t} = -r^{\text{HVA4}} + g^{\text{HVA4}} \cdot \frac{E \cdot A}{\sigma + S} \tag{27}$$

$$S_{d,i,x} = E \cdot (A + S^{\text{FEAT}} + S^{\text{SP}} + S^{\text{SUR}}) \tag{28}$$

Whereby E denotes excitation, A amplification, and S suppression from an associated interneuron. The interneuron receives several sources of suppression: the excitation, feature-based suppression from dissimilar features in HVA layer 2/3 at the same location (S^{FEAT}), spatial suppression from FEFvm at all other locations (S^{SP}), and surround suppression from similar features in the surround of HVA layer 2/3 (S^{SUR}). The parameter $\tau^{\text{HVA4}} = 10$ denotes the time constant, $\sigma = 0.4$ the attention contrast gain factor, and $g^{\text{HVA4}} = 1.066$ an factor to reach a maximal response of 1 (similar to R_{Max} in Albrecht and Hamilton (1982)).

HVA is able to represent different visual stimuli to adapt the model to the needs of a specific application scenario. In object localization, the area represents object views. Both scenarios utilize learned representations. Yet, learning is not even necessary in many psychophysical setups as they use very simple stimuli that can be represented via simple features like color or orientation. Such simple features are encoded already in V1. Hence, we implement in our model the possibility to represent the same features also in HVA. In this case, HVA contains the same three channels as V1 ($d = \{1, 2, 3\}$), whereby it contains a single channel in case of learned representations ($d = \{1\}$).

The excitation to a HVA layer 4 cell is received from complex cells in V1, and is either calculated via learned descriptors (Eq. 29), or via pooling of V1 features (Eq. 30). Both variations can be scaled via $v^{\text{V1-HVA4}} = 1$ and a non-linearity $p^{\text{E}} = 1$. The connectivity weights $w^{\text{V1-HVA4}}$ are either provided by an external learning procedure, or are modeled by a 2D-Gaussian (Eq. 31).

$$E_{d,i,x} = \left[v^{\text{V1-HVA4}} \cdot f_2 \left(\sum_{d',i',x' \in \text{RF}(x,V1)} \left(w^{\text{V1-HVA4}}_{i,\,d',i',x'}\, r^{\text{V1C}}_{d',i',x'} \right) \right) \right]^{p_E} \tag{29}$$

$$E_{d,i,x} = \left[v^{\text{V1-HVA4}} \cdot f_2 \left(\max_{x' \in \text{RF}(x,V1)} \left(w^{\text{V1-HVA4}}_{x'}\, r^{\text{V1C}}_{d,i,x'} \right) \right) \right]^{p_E} \tag{30}$$

$$w^{\text{V1-HVA4}}_{x'} = \mathfrak{g}(x', 1, [8.3, 8.3]) \tag{31}$$

A cell receives spatial amplification from the same location in FEFvm (Eq. 33), and feature-based amplification from the same feature in HVA layer 2/3 (Eq. 34). The spatial amplification is modeled via a one-to-one connection with the scaling parameter $v^{\text{FEFvm-HVA4}} = 4$. The connection from the FEF to layer 4 is inspired by the anatomical finding that the FEF targets layer 4 in the visual area V4 (Barone et al., 2000). The feature-based amplification is modeled via a 2D-Gaussian connectivity ($w^{\text{HVA2-HVA4}}$, Eq. 35), reciprocally to the feedforward connections from HVA layer 4 to layer 2/3. The amplification can be tuned by a scaling parameter $v^{\text{HVA2-HVA4}} = 1$ and a non-linearity parameter $p^{\text{HVA2-HVA4}} = 1$. The effects of spatial and feature-based amplification are summed up additively (Eq. 32) as multiple studies show an additive influences of both attention forms (Saenz et al., 2002; Treue and Trujillo, 1999).

$$A_{d,i,x} = 1 + A^{\text{SP}} + A^{\text{FEAT-L4}} \tag{32}$$

$$A^{\text{SP}}_{d,i,x} = v^{\text{FEFvm-HVA4}}\, \bar{r}^{\text{FEFvm}}_{x} \tag{33}$$

$$A^{\text{FEAT-L4}}_{d,i,x} = v^{\text{HVA2-HVA4}} \left(\max_{x' \in \text{RF}(x,HVA2)} w^{\text{HVA2-HVA4}}_{x'}\, r^{\text{HVA2}}_{d,i,x'} \right)^{p_{HVA2-HVA4}} \tag{34}$$

$$w^{\text{HVA2-HVA4}}_{x'} = \mathfrak{g}(x', 1, [0.6, 0.6]) \tag{35}$$

Feature-based suppression is received from neurons in layer 2/3 preferring a dissimilar feature. The connectivity matrix w^{FEAT} differs dependent on whether HVA encodes simple features or learned object view descriptors. For the former case, the weights are channel-specific squared functions (Eq. 36). For the latter case, the feature-based suppression inhibits only view cells belonging to different objects, thus w is zero between cells belonging to the same object (Eq. 37). The strength is normalized with the number of view cells encoding a particular object. View cells encoding the same object k are typically simultaneously active, thus there exist $\#L_k$ pre-synaptic cells i' which simultaneously suppress a post-synaptic cell i. A normalization with $1/\#L_k$ ensures that the post-synaptic cell i will receive the same amount of suppression, independently of the number of pre-synaptic cells $\#L_k$ (Eq. 37).

If HVA encodes simple features:

$$w^{\text{FEAT}}_{d,i,i'} = \begin{cases} (|i-i'|/7)^2 & d = 1,2 \\ (|i-i'|/3)^2 & d = 3, |i-i'| \leq 3 \\ 1 - (|i-i'|-4)/3)^2 & d = 3, |i-i'| > 3 \end{cases} \tag{36}$$

If HVA encodes object views:

$$w^{\text{FEAT}}_{d,i,i'} = \begin{cases} 0 & \text{Views cells } i \text{ and } i' \text{ belong to the same object.} \\ 1/\#L_k & \text{Else.} \end{cases} \tag{37}$$

The feature-based suppression (Eq. 40) is received from all locations in layer 2/3 as it is functional necessary to suppress objects between all locations in the localization task. However, this is in contrast to the microcircuit model (Beuth and Hamker, 2015) which proposes feature-based suppression only within a local region, similar in size as the receptive field. The earlier model of Hamker (2005b) shows how the brain might implement long-range suppression under consideration of this constraint. It realizes the suppression via a connection chain over two high-level visual areas, V4 and IT. Area IT has very large receptive fields and thus can contains also long-range suppressive connections. We model our area HVA to properties of both areas, thus we abstract also this connectivity chain to one connection and model directly long-range suppressive connections. However, if HVA encodes simple features, the suppression (Eq. 39) is received from the local region as proposed by the microcircuit model. The connectivity $w^{\text{HVA2-HVA4}}$ is simulated by the same Gaussian function (Eq. 35) as the feature-based amplification.

Non-linearities are implemented via power-functions (x^p) to receive a greater amount of suppression from highly active neurons (parameter $p^{\text{FEAT-2}} = 2$), as well as to scale the total influence of the suppression non-linearly ($p^{\text{FEAT-1}} = 3$). The former is especially necessary for network configuration with a larger number of view cells. We use configurations with 20, 80, and 354 view cells to represent the three different object sets with 5, 15, and 100 objects. The parameter $p^{\text{FEAT-2}}$ is set to 2 for 20 cells, to 3 for 80 cells, and to 5 for 354 cells. The two non-linearities allow, together with two scaling factors $v^{\text{FEAT-1}} = 3$ and $v^{\text{FEAT-2}} = 2$, a fine graded tuning of the suppression.

$$S^{\text{FEAT}}_{d,i,x} = \left[v^{\text{FEAT-1}} \cdot f_2 \left(\sum_{i'} w^{\text{FEAT}}_{d,i,i'} \cdot \left(v^{\text{FEAT-2}}\, B_{d,i',x} \right)^{pFEAT-2} \right) \right]^{pFEAT-1} \tag{38}$$

If HVA encodes simple features:

$$B_{d,i,x} = \max_{x' \in \text{RF}(x,HVA2)} \left(w^{\text{HVA2-HVA4}}_{x'}\, r^{\text{HVA2}}_{d,i,x'} \right) \tag{39}$$

If HVA encodes object views:

$$B_{d,i,x} = \max_{x'} \left(r^{\text{HVA2}}_{d,i,x'} \right) \tag{40}$$

Spatial suppression is received from other retinotopic locations in the FEFvm (Eq. 41). Its parameters control again separately the non-linear influence of pre-synaptic neurons ($v^{\text{SP-2}} = 1$, $p^{\text{SP-2}} = 1$), and the non-linear influence of the total connection ($v^{\text{SP-1}} = 0.85$, $p^{\text{SP-1}} = 1$). The weight matrix w^{SP} (Eq. 42) models a long-range inhibition which diminish to zero at close locations by a sharp negative 2D-Gaussian. We choose a long-range inhibition due to psychophysical evidences and functional requirements. The psychophysical study of Caputo and Guerra (1998) investigate distractor suppression. They found a strong surround suppression and an average-strong long-range suppression. Our model contains already a surround suppression (next paragraph), thus we model here only the long-range suppression aspect. Furthermore, we found that the long-range inhibition is required for the function of the model. The spatial processing in the

FEF centers neural activity at the target location and suppresses the location of distractors during this process. We found that it is necessary to suppress the distractors in HVA too, and we realize this via such inhibitory connections.

$$S_x^{\text{SP}} = \left[v^{\text{SP-1}} \cdot \sum_{x' \in \text{RF}(x, FEFvm)} w_{x'}^{\text{SP}} \cdot (v^{\text{SP-2}}\, \bar{r}_{x'}^{\text{FEFvm}})^{p_{SP-2}} \right]^{p_{SP-1}} \tag{41}$$

$$w_{x'}^{\text{SP}} = \left(1 - 2\,(\mathfrak{g}(x', 1, [3,4])^{0.125})\right)^{+} \tag{42}$$

Surround suppression is received from layer 2/3 neurons encoding the same feature at surround locations (Eq. 44). This kind of suppression is disables by default ($v^{\text{SUR-1}} = 0$) as it is a not necessary in the object localization setup. Nevertheless, we include it in the system-level model as it is a part of the microcircuit model. The parameters control again separately the influence of presynaptic neurons ($v^{\text{SUR-2}} = 2$, $p^{\text{SUR-2}} = 2$), and of the total connection ($v^{\text{SUR-1}} = 0$, $p^{\text{SUR-1}} = 1$). The connectivity w^{SUR} is shaped as a ring. We use for this purpose the surround part of a difference of 2D-Gaussians (Eq. 43).

$$w_{x'}^{\text{SUR}} = \frac{K}{|K|}, \text{ with: } K = \left(\mathfrak{g}(x', 1, [6,6]) - \mathfrak{g}(x', 2, [3,3])\right)^{+} \tag{43}$$

$$S_{d,i,x}^{\text{SUR}} = \left[v^{\text{SUR-1}} \cdot \sum_{x' \in \text{RF}(x, HVA2)} w_{x'}^{\text{SUR}} \cdot \left(v^{\text{SUR-2}}\, r_{d,i,x'}^{\text{HVA2}}\right)^{p_{SUR-2}} \right]^{p_{SUR-1}} \tag{44}$$

Higher visual area - layer 2/3 HVA layer 2/3 pools spatially layer 4 responses (Eq. 45, 46), whereby the pooling is executed for each feature and channel separately .

$$\tau^{\text{HVA2}} \frac{\partial r_{d,i,x}^{\text{HVA2}}}{\partial t} = -r^{\text{HVA2}} + g^{\text{HVA2}} \cdot \frac{E \cdot (1 + A^{\text{FEAT-L2}})}{\sigma + S} \tag{45}$$

$$S_{d,i,x} = E \cdot (1 + A^{\text{FEAT-L2}}) \tag{46}$$

Whereby E denotes excitation, A feature-based amplification, and S suppression. The parameter $\tau^{\text{HVA2}} = 10$ denotes the time constant, $\sigma = 1$ the attention contrast gain factor, and $g^{\text{HVA4}} = 1.69$ a factor to reach a maximal response of 1 (similar to R_{Max} in Albrecht and Hamilton (1982)).

Excitation results from pooling layer 4 features (Eq. 47), whereby $v^{\text{HVA4}} = 1$ controls its influence. The pooling is implemented via a soft-max (Beuth and Hamker, 2015), whereby $p_1 = 4$ and $p_2 = 0.25$ parametrize the involved nonlinearities. The connectivity is implemented via a 2D-Gaussian, thus a layer 2/3 cell reacts less powerful at the borders of their receptive field (Heuer and Britten, 2002). A Gaussian-modulated pooling has been used also in previous models, e.g. Hamker (2005a,b); Hamker and Zirnsak (2006)

$$E_{d,i,x} = \left(v^{\text{HVA4}} \cdot \sum_{x' \in \text{RF}(x, HVA4-HVA2)} w_{x'}^{\text{HVA4-HVA2}}\, (r_{d,i,x'}^{\text{HVA4}})^{p_1} \right)^{p_2} \tag{47}$$

$$w_{x'}^{\text{HVA4-HVA2}} = \mathfrak{g}(x', 1, [1,1]) \tag{48}$$

Feature-based amplification is received from the prefrontal cortex (PFC, Eq. 49), whereby the parameter $v^{\text{PFC-HVA2}} = 1.5$ scales its influence. If HVA encodes object views, amplification is received from the associated object via an external learned connectivity matrix $w^{\text{PFC-HVA2}}$. Otherwise, it is received from the same feature via a one-to-one connection, thus $w^{\text{PFC-HVA2}}$ is the identity matrix. The amplification signal represents the top-down attentional influence of the PFC on high-level visual areas like IT (Buschman and Miller, 2007; Fuster, 2000; Tomita et al., 1999). The signal is part of the attentional processing network in the cortex (Miller and Buschman, 2013). However, it is currently unsettled which connectivity transports this signal. We assume the simplest option of a direct connection as at least the reverse connection from IT to the PFC exists (Barbas, 2000; Seltzer and Pandya, 1989). Yet, other possibilities are a transmission via the pulvinar (Draganski et al., 2008; Steele and Weller, 1993), or via the medial temporal lobe (Miller and Cohen, 2001; Kravitz et al., 2013).

$$A^{\text{FEAT-L2}}_{d,i,x} = v^{\text{PFC-HVA2}} \cdot \sum_{i'} w^{\text{PFC-HVA2}}_{d,i,i'} \, r^{\text{PFC}}_{d,i'} \tag{49}$$

Prefrontal cortex (PFC) A simple model of the prefrontal cortex (PFC) encodes cells reacting to specific object categories, i.e. to an object under all view-points. These cells simulate the object-category specific cells of the primate prefrontal cortex (Ashby and Spiering, 2004; Freedman et al., 2001; Seger and Miller, 2010). Yet, the PFC is also involved in many other high-level functions (Miller and Cohen, 2001; Seger, 2008; Vitay and Hamker, 2010).

To meet different task demands, our PFC is is either able to store the category of the currently presented stimulus (recognition task), or to encode the search target when searching for a particular object (localization task). A parameter *PFCtarget* stores the search target. If the parameter is defined, the model performs a localization task, and otherwise a recognition task.

The PFC can encode varying entities dependent on the mapping from HVA layer 2/3 to PFC. In the object localization task, we used a learned mapping to encodes object categories (Sec. 2). We simulate one cell for each object i: r^{PFC}_i. If a learned mapping is not necessary, a simple one-to-one connection can be used. The PFC would then encode the same features i and channels d as HVA layer 2/3: $r^{\text{PFC}}_{d,i}$. This configuration could be used for simple psychophysical setups in which HVA encodes simple features.

Frontal eye field - visual cells (FEFv) The frontal eye field (FEF) processes spatial information, selects the target location, and controls eye movements (Heinzle, 2006; Pouget et al., 2009). Our model of the FEF is based on Zirnsak et al. (2011). It has been developed for simple psychophysical stimuli in a single scene and we found that it is not able to deal with strongly varying scene statistics as in the object localization task. For example, those scenes differ in the amount of background clutter, the saliency of the objects, the saliency of the distractors, etc. Thus, we modernize the underlying equations under consideration of the originally modeled physiological properties. Therefore, we would refer the reader to the original publication for physiological background and focus here on functional aspects.

We model four cell types of the FEF (Schall, 1991): visual (FEFv), visuomovement (FEFvm), movement (FEFm), and fixation cell types (FEFfix). The FEFv simulate the visual cells (Segraves and Goldberg, 1987; Schall, 1991). The map receives inputs from visual cortices at the same retinotopic location, irrespective of the feature information and thus, encodes the visual conspicuities (Hamker, 2005a). This representation is often denoted as saliency map (Itti and Koch, 2001).

The FEFv is excited from HVA layer 2/3 (Eq. 53) and from V1 (Eq. 54). HVA layer 2/3 projects to the FEFv via a one-to-one connection, and V1 to FEFv via a 2D-Gaussian connectivity matrix (Eq. 55). The latter simulates the fast dorsal pathway LGN→MT→FEF in the cortex (Heinzle, 2006; Sincich et al., 2004). The connection pools V1 responses spatially, which is similar implemented as in HVA layer 2/3 via non-linearities ($p_1 = 2$, $p_2 = 0.5$). The pathway is not necessary in the object localization scenario and hence disabled ($v^{\text{V1-FEFv}} = 0$). Nevertheless, we include it also for the generality of the model. The robustness of the FEF was improved regarding varying scene statistics by a non-linearity C to increase the difference between weak and strong input signals (Eq. 56) (Antonelli et al., 2014), and by a signal enhancement operation Q based on divisive normalization (Eq. 57). The latter decouples the effects of feature-based suppression from the spatial processing in the FEF. The FEF response remains strong even if the response in HVA layer 2/3 is suppressed.

$$\tau^{\text{FEFv}} \frac{\partial r_x^{\text{FEFv}}}{\partial t} = -r^{\text{FEFv}} + E \tag{50}$$

$$E_x = C\left(Q\left(F_x\right)\right) \tag{51}$$

$$F_x = \max\left\{E_x^{\text{V1}}, E_x^{\text{HVA2}}\right\} \tag{52}$$

$$E_x^{\text{HVA2}} = \max_{d',i'}\left(r_{d',i',x}^{\text{HVA2}}\right) \tag{53}$$

$$E_x^{\text{V1}} = v^{\text{V1-FEFv}} \cdot \max_{d',i'}\left(\left(\sum_{x' \in \text{RF}(x,FEFv)} w_{x'}^{\text{V1-FEFv}} \left(r_{d',i',x'}^{\text{V1C}}\right)^{p_1}\right)^{p_2}\right) \tag{54}$$

$$w_{x'}^{\text{V1-FEFv}} = \mathfrak{g}(x', 1, [36.6, 36.6]) \tag{55}$$

$$C(x) = (x \cdot (1 + c) - c)^{+} \tag{56}$$

$$Q(x) = x \cdot \frac{(1 + \sigma)}{F^{\max} + \sigma} \tag{57}$$

Whereby E denotes the excitation, $\tau^{\text{FEFv}} = 10$ the time constant, $c = 6$ a competition parameter, and $F^{\max}$ describes the maximum over all F_x.

Frontal eye field - visual movement cells (FEFvm) The visuomovement cells in the FEF react to visual stimuli, but also encode saccade target information (Ray et al., 2009). Our modeled cells encode a continuous spectrum of both influences (Zirnsak et al., 2011). The visual information is transmitted via inputs from FEFv (E^{FEFv} in Eq. 60), and the saccadic information via inputs from FEFm (r^{FEFm} in Eq. 58). Their influences are proportionally weighted to each other via $v_i^{\text{FEFv-FEFvm}}$ and $1 - v_i^{\text{FEFv-FEFvm}}$ (Eq. 58).

$$\tau^{\text{FEFvm}} \frac{\partial r_{i,x}^{\text{FEFvm}}}{\partial t} = -r_{i,x}^{\text{FEFvm}} + v_i^{\text{FEFv-FEFvm}}\, E^{\text{FEFv}} + \left(1 - v_i^{\text{FEFv-FEFvm}}\right) r_x^{\text{FEFm}} \tag{58}$$
$$\text{with:} \quad \tau^{\text{FEFvm}} = 10$$

The input signal from FEFv realizes a competition between locations via a local Gaussian excitation (Eq. 60) and a long-range suppression (Eq. 61). Such a long-range, spatial competition is typically employ by models of visual search, despite its precise mechanisms are unknown. We realize here the popular idea that the competition is mediated by inhibitory connections within the FEF (Pouget et al., 2009). The suppression can be fine-tuned via a non-linearity parameter $p^{\text{Sv-1}} = 1$ and scaling parameters $v^{\text{Sv-1}} = 0.6$, $v^{\text{Sv-2}} = 0.35$ (Eq. 61). The excitation can be scaled similarly via $v^{\text{Ev}} = 0.6$ (Eq. 60). The competition is a part of the recurrent processing within the loop HVA layer 2/3 → FEFv →FEFvm → HVA layer 4 → HVA layer 2/3. We presume that the suppression cannot completely inhibit the visuomovement cells, thus they are always weakly-driven by visual stimuli. This assumption prevents a disruption of the spatial processing loop as it avoids an extinction of neuronal activity in case of a strong suppression from FEFv. Strong suppression typically occurs in crowded scenes as they evoke a very broad response in the FEFv, resulting in a strong suppression of the FEFvm. The weak visual excitation is implemented by a Gaussian and a ratio factor $v^{\text{low}} = 0.2$ (Eq. 59).

$$E_x^{\text{FEFv}} = v^{\text{low}} \cdot f_1\left(E\right) + \left(1 - v^{\text{low}}\right) \cdot f_2\left(E - S\right) \tag{59}$$
$$E_x = v^{\text{Ev}} \cdot \sum_{x'} w_{x'}^{\text{Ev}}\, r_{x'}^{\text{FEFv}} \tag{60}$$
$$S_x = \left(v^{\text{Sv-1}} \cdot \sum_{x'} w_{x'}^{\text{Sv}}\, r_{x'}^{\text{FEFv}}\right)^{p_{Sv-1}} \tag{61}$$
$$K_{x'} = \mathfrak{g}(x', 1, [3,4]) - v^{\text{Sv-2}} \tag{62}$$
$$w_{x'}^{\text{Ev}} = (K)^{+} \tag{63}$$
$$w_{x'}^{\text{Sv}} = (-K)^{+} \tag{64}$$

Frontal eye field - movement (FEFm) and fixation cells (FEFfix) The FEFm represent the movement cells of the frontal eye field (Bruce and Goldberg, 1985; Segraves and Goldberg, 1987). They encode eye movement related information, for example their activity ramps up shortly before the execution of a saccade. This finding leads to the proposal that they encode the target location of a planned saccade (Hamker, 2005b). FEFm (Eq. 65) focuses neuronal activity to a single saccade target location by a competition among locations. The competition is implemented via the input signal from FEFvm by a local, point-wise excitation (Eq. 66) and a long range inhibition (Eq. 67). Their influences can be calibrated by the parameters $v^{\text{FEFvm-FEFm}} = 1.3$ and $v^{\text{Svm}} = 0.3$.

The FEFfix represents the fixation cells of the FEF (Hasegawa et al., 2004; Hamker, 2005b) that suppresses the execution of saccades. Their precise mechanisms are unsettled, thus we model the simplest approach of a single cell suppressing globally saccades (r^{FEFfix}, Eq. 68). The cell's influence can be tuned via the parameter $v^{\text{Sfix}} = 3$. Suppression of saccades depends typically on the experimental setup, hence the cell activity should be set by the user.

$$\tau^{\text{FEFm}} \frac{\partial r_x^{\text{FEFm}}}{\partial t} = -r_x^{\text{FEFm}} + E_x^{\text{FEFvm}} - S_x^{\text{FEFvm}} - S^{\text{Fix}}, \quad \text{with: } \tau^{\text{FEFm}} = 10 \tag{65}$$

$$E_x^{\text{FEFvm}} = v^{\text{FEFvm-FEFm}}\, \overline{r}^{\text{FEFvm}} \tag{66}$$

$$S_x^{\text{FEFvm}} = v^{\text{Svm}} \max_{x'} \overline{r}_{x'}^{\text{FEFvm}} \tag{67}$$

$$S^{\text{FEFfix}} = v^{\text{Sfix}}\, r^{\text{FEFfix}} \tag{68}$$

If the response exceeds a threshold Γ^{FEFm} at time point t_o, we assume the upcoming execution of a saccade. The threshold idea (Hamker, 2005b) is inspired from the finding that a saccade is executed when the movement-related FEF activity reaches a certain level (Hanes and Schall, 1996). The saccade target location is calculated by the center of gravity from the FEFm activity (Eq. 69).

$$x_c = \frac{\sum_{x'} r_{x'}^{\text{FEFm}}(t_o) \cdot x'}{\sum_{x'} r_{x'}^{\text{FEFm}}(t_o)} \tag{69}$$

2 Learning of object representations

The attention model performs the object localization task by a neuronal view-based representation of all objects. We choose objects from the COIL-100 database (Nene et al., 1996) in three sets with 5, 15 and 100 objects and generate a representation for each one. The representation were created in an external, offline-learning stage and afterwards loaded into the model. Thus, they are fixed during the model execution. We employ an unsupervised, trace learning approach relaying on temporal continuity (Földiák, 1990; Spratling, 2005; Teichmann et al., 2012), which leads to partly rotation invariant representation of an object view. The idea is that on the short time scale of stimuli presentations, the visual input is more likely to originate from the same object, rather than from different objects. The algorithm must be trained on an image sequence resembling the temporal behavior of the retinal image stream. Our trace learning assumes that on average, rotations of the target object and saccades in its vicinity are more likely than saccades to different objects. Therefore, the sequence of images was arranged such that the object view changes often (every 50 ms by 10 degree), and the objects type rarely (randomly every 5.4 s). The training was performed on the image sequence until all object representations were stable, which requires about 100 000 presented images or 5000 s simulation time. From this sequence, the algorithm learns a population of view-tuned HVA cells that encode the statistically significant information of a certain view point of an object, hence they react to a specific view of an object (Fig. 2a). The population of view-tuned cells was learned on a single HVA location and then shared with all other locations (weight sharing).

The trace learning is supported by Anti-Hebbian learning (Antonelli et al., 2014; Beuth et al., 2010; Wiltschut and Hamker, 2009) to increase the competition among objects. Competition is strengthened when HVA features are activated simultaneously. The Anti-Hebbian learning leads to decorrelated responses and to a sparse code of the neuronal population (Földiák, 1990).

The learning algorithm is similar to Antonelli et al. (2014) and thus we illustrate only its core component (Eq. 70). The learning equation associates the HVA response r_i^{HVA} of the previous stimulus at $t-1$ with the V1 representation $r_{i'}^{\text{V1}}$ of the current stimulus. The HVA response of the previous stimulus acts as a neuronal trace and so realizes the temporal continuity learning (Spratling, 2005).

$$\begin{aligned}\tau_w \frac{\partial w_{i',i}^{\text{V1-HVA}}}{\partial t} &= (r_{i'}^{\text{V1}} - \theta^{\text{V1}})_t \cdot (r_i^{\text{HVA}} - \theta^{\text{HVA}})_{t-1}^{+} \\ &\quad - \alpha_w \cdot w_{i',i} \cdot (r_i^{\text{HVA}} - \theta^{\text{HVA}})_{t-1}^{2} \\ \text{with: } \alpha_w &= \begin{cases} \alpha_{w+} & w_{i',i}^{\text{V1-HVA}} \geq 0 \\ \alpha_{w-} & w_{i',i}^{\text{V1-HVA}} < 0 \end{cases}\end{aligned} \tag{70}$$

Whereby $\alpha_{w+} = 80$ and $\alpha_{w-} = 20$ constrains the weights, $\tau_w = 10^4$ is a time constant controlling the speed of the learning process, and $[x]^+$ stands for argmax$(x, 0)$. The term $\theta^{\text{V1}} = \bar{r}^{\text{V1}}$ is the mean activation of the whole population of V1, while $\theta^{\text{HVA}} = \max(\gamma \cdot \max(r^{\text{HVA}}), \bar{r}^{\text{HVA}})$ with $\gamma = 0.9$. We use three different learning configurations to generate the three objects sets. The outlined values describe the configuration with 5 objects, whereby the values for 15 objects are: $\alpha_{w+} = 60,\ \alpha_{w-} = 7.5,\ \tau_w = 250,\ \gamma = 0.975$; and for 100 objects: $\alpha_{w+} = 60,\ \alpha_{w-} = 7.5,\ \tau_w = 400,\ \gamma = 0.975$.

The mapping from objects to views is represented by the connections between PFC and HVA (Fig. 2b). They must be learned due to the large number of objects, which is a necessary improvement to the manually designed weights in previous work (Antonelli et al., 2014; Beuth et al., 2010). The mapping is used to send the amplification signal to all HVA cells belonging to the target object. A supervised learning procedure determines the HVA cells belonging to each object. After the trace learning, we present the training stimuli plus their object IDs to the learning network, and record which HVA cells respond strongly for a particular object. The connection strength is set to 1 for these combinations, and to 0 otherwise. A HVA cell i is defined as strongly responding if its response r_i^{HVA} is over the threshold θ_{HVA} (Eq. 70). Only such a cell has learned the presented object during the trace learning as Eq. 70 is zero for $r_i^{\text{HVA}} < \theta_{HVA}$. Thus, we connect precisely the HVA cells that encode the presented object.

If a HVA cell is associated to multiple objects (red marked in Fig. 2b), we keep only a single association to ensure a top-down signal specifically targeting a

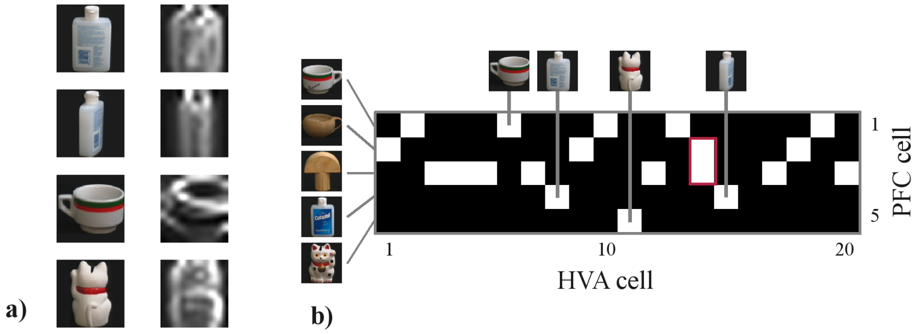

Figure 2. a) Encoding of objects via learned view-tuned cells. For each object view (left), the weights V1→HVA of one exemplary HVA cell (right) are illustrated as the maximum over all V1 features and channels. Brightness denotes weight strength. **b)** The mapping from object cells in PFC to view cells in HVA, after learning a set of five objects. Connected cells are indicated by white color. The examples from (a) are illustrated, too. The red rectangle denotes a HVA cell which is incorrectly associated to multiple objects.

single object. Typically, such a cell reacts to many trainings stimuli of an object A, but also for a few stimuli of another object B. As most of the stimuli belong to object A, it indicates that the cell encodes object A. Thus, we solely connect the cell to object A. If we would connect the view-tuned cell also to object B, it would impair the search if object B is the target and the scene contains both objects. The task implies to amplify all HVA cells belonging to object B, but as these cells react primary to object A, they will react more at the location of object A than of B, and the model would incorrectly select object A.

Bibliography

Albrecht, D. and Hamilton, D. (1982). Striate cortex of monkey and cat: Contrast response function. *J Neurophysiol*, 48(1):217–37.

Antonelli, M., Gibaldi, A., Beuth, F., Duran, A. J., Canessa, A., Chessa, M., Hamker, F. H., Chinellato, E., and Sabatini, S. P. (2014). A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. *IEEE Trans Auton Mental Develop*, 6(4):259–273.

Ashby, F. G. and Spiering, B. J. (2004). The neurobiology of category learning. *Behav Cogn Neurosci Rev*, 3(2):101–13.

Atkinson, K. (1989). *An introduction to numerical analysis.*

Barbas, H. (2000). Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. *Brain Res Bull*, 52(5):319–30.

Barone, P., Batardiere, A., Knoblauch, K., and Kennedy, H. (2000). Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hierarchical rank and indicates the operation of a distance rule. *J Neurosci*, 20(9):3263–81.

Beuth, F. and Hamker, F. H. (2015). A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Res.*

Beuth, F., Wiltschut, J., and Hamker, F. H. (2010). Attentive Stereoscopic Object Recognition. In Villmann, T. and Schleif, F.-M., editors, *Proc Workshop New Challenges in Neural Computation 2010 - NCNC 2010, Machine Learning reports 04/2010, AG Computational Intelligence, University of Leipzig*, pages 41–48.

Bowmaker, J. and Dartnall, H. (1980). Visual Pigments of Rods and Cones in a Human Retina. *J Physiol*, 298:501–511.

Bruce, C. J. and Goldberg, M. E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. *J Neurophysiol*, 53(3):603–35.

Buschman, T. J. and Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–2.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., and Poggio, T. (2007). A model of V4 shape selectivity and invariance. *J Neurophysiol*, 98(3):1733–50.

Caputo, G. and Guerra, S. (1998). Attentional selection by distractor suppression. *Vision Res*, 38(5):669–89.

Chatterjee, S. and Callaway, E. M. (2003). Parallel colour-opponent pathways to primary visual cortex. *Nature*, 426:668–71.

Dacey, D. M. (2000). Parallel pathways for spectral coding in primate retina. *Annu Rev Neurosci*, 23:743–775.

Draganski, B., Kherif, F., Klöppel, S., Cook, P. a., Alexander, D. C., Parker, G. J. M., Deichmann, R., Ashburner, J., and Frackowiak, R. S. J. (2008). Evidence for segregated and integrative connectivity patterns in the human Basal Ganglia. *J Neurosci*, 28(28):7143–52.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol Cybern*, 237(5349):55–56.

Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312.

Fuster, J. M. (2000). Executive frontal functions. *Exp Brain Res*, 133(1):66–70.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Rev Neurosci*, 4(7):563–72.

Getreuer, P. (2010). MATLAB function colorspace.m.

Hamker, F. H. (2005a). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Comput Vis Image Underst*, 100:64–106.

Hamker, F. H. (2005b). The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex*, 15(4):431–47.

Hamker, F. H. and Zirnsak, M. (2006). V4 receptive field dynamics as predicted by a systems-level model of visual attention using feedback from the frontal eye field. *Neural Netw*, 19(9):1371–82.

Hanes, D. and Schall, J. (1996). Neural control of voluntary movement initiation. *Science*, 274(5286):427–430.

Hasegawa, R. P., Peterson, B. W., and Goldberg, M. E. (2004). Prefrontal neurons coding suppression of specific saccades. *Neuron*, 43(3):415–25.

Hegdé, J. and Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas v2 and v4. *Cerebral cortex*, 17(5):1100–16.

Heinzle, J. (2006). *A model of the local cortical circuit of the frontal eye fields*. PhD thesis.

Heuer, H. W. and Britten, K. H. (2002). Contrast dependence of response normalization in area MT of the rhesus macaque. *J Neurophysiol*, 88(6):3398–408.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Rev Neurosci*, 2:194–203.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol*, 58(6):1233–58.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cogn Sci*, 17(1):26–49.

Kriegeskorte, N. (2009). Relating Population-Code Representations between Man, Monkey, and Computational Models. *Front Neurosci*, 3(3):363–73.

Logothetis, N. K., Pauls, J., and Poggiot, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr Bio*, 5(5):552–563.

Miller, E. K. and Buschman, T. J. (2013). Cortical circuits for the control of attention. *Curr Opin Neurobiol*, 23(2):216–222.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, 24(1):167–202.

Moroney, N., Fairchild, M. D., Hunt, R. W. G., Li, C., Luo, M. R., Newman, T., Laboratories, H.-p., Alto, P., Color, M., Consultant, C., Americas, C. D., and Jose, S. (2002). The CIECAM02 Color Appearance Model. In *Proc IS&T/SID 10th Color Imaging Conf. - CIC 2002*, pages 23–27.

Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia Object Image Library (COIL-100). *Technical Report CUCS-006-96.*

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci*, 4(12):1244–52.

Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area V4. *Nat Neurosci*, 5(12):1332–38.

Pouget, P., Stepniewska, I., Crowder, E. a., Leslie, M. W., Emeric, E. E., Nelson, M. J., and Schall, J. D. (2009). Visual and motor connectivity and the distribution of calcium-binding proteins in macaque frontal eye field: implications for saccade target selection. *Front Neuroanat*, 3:2.

Ray, S., Pouget, P., and Schall, J. D. (2009). Functional distinction between visuomovement and movement neurons in macaque frontal eye field during saccade countermanding. *J Neurophysiol*, 102:3091–3100.

Saenz, M., Buracas, G. T., and Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nat Neurosci*, 5(7):631–32.

Schall, J. D. (1991). Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields. *J Neurophysiol*, 66(2):559–79.

Seger, C. a. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci Biobehav Rev*, 32(2):265–78.

Seger, C. a. and Miller, E. K. (2010). Category learning in the brain. *Annu Rev Neurosci*, 33:203–19.

Segraves, A. and Goldberg, E. (1987). Functional properties of corticotectal neurons in the monkey's frontal eye field. *J Neurophysiol*, 58(6):1387–1419.

Seltzer, B. and Pandya, D. (1989). Frontal Lobe Connections of the Superior Temporal Sulcus. *J Comp Neurol*, 281:97–113.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007). A quantitative theory of immediate visual recognition. In Cisek, P., Drew, T., and Kalaska, J., editors, *Prog Brain Res*, volume 165, pages 33–56.

Sincich, L. C. and Horton, J. C. (2005). The circuitry of V1 and V2: integration of color, form, and motion. *Annu Rev Neurosci*, 28:303–26.

Sincich, L. C., Park, K. F., Wohlgemuth, M. J., and Horton, J. C. (2004). Bypassing V1: a direct geniculate input to area MT. *Nat Neurosci*, 7(10):1123–8.

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

Spratling, M. W. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Trans Pattern Anal Mach Intell*, 27(5):753–61.

Steele, G. E. and Weller, R. E. (1993). Subcortical connections of subdivisions of inferior temporal cortex in squirrel monkeys. *Vis Neurosci*, 10(3):563–583.

Stockman, a. and Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Res*, 40(13):1711–37.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci*, 19:109–39.

Teichmann, M., Wiltschut, J., and Hamker, F. H. (2012). Learning invariance from natural images inspired by observations in the primary visual cortex. *Neural Comput*, 24(5):1271–96.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., and Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, 401(6754):699–703.

Treue, S. and Trujillo, J. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–579.

Turkowski, K. and Gabriel, S. (1990). Filters for common resampling tasks. In *Graphics gems*, pages 147–165.

Vitay, J. and Hamker, F. H. (2010). A computational model of Basal Ganglia and its role in memory retrieval in rewarded visual memory tasks. *Front Comp Neurosci*, 4:13.

Wiesel, T. and Hubel, D. (1966). Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J Neurophysiol*, 29(6):1115–56.

Wiltschut, J. and Hamker, F. H. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Vis Neurosci*, 26(1):21–34.

Zirnsak, M., Beuth, F., and Hamker, F. H. (2011). Split of spatial attention as predicted by a systems-level model of visual attention. *Eur J Neurosci*, 33(11):2035–45.