Frederik Beuth

Visual attention in primates and for machines - neuronal mechanisms

Frederik Beuth

Visual attention in primates and for machines - neuronal mechanisms

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Universitätsverlag Chemnitz
2019

Impressum

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Angaben sind im Internet über http://www.dnb.de abrufbar.

Titelgrafik: Frederik Beuth, Shabnam Novin
Satz/Layout: Frederik Beuth

Technische Universität Chemnitz/Universitätsbibliothek
Universitätsverlag Chemnitz
09107 Chemnitz
https://www.tu-chemnitz.de/ub/univerlag

readbox unipress
in der readbox publishing GmbH
Am Hawerkamp 31
48155 Münster
http://unipress.readbox.net

ISBN 978-3-96100-102-6

https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-356553

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Professorship Artificial Intelligence

Visual attention in primates and for machines - neuronal mechanisms

Dissertation

submitted for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

in the

Department of Computer Science,
Technische Universität Chemnitz

by Frederik Beuth

Day of the submission: December 19, 2017
Day of the defense: October 30, 2018

Examiner: Prof. Dr. Fred H. Hamker, Technische Universität Chemnitz
Prof. Dr. Heiko Neumann, Universität Ulm
Prof. Dr. Wolfgang Einhäuser-Treyer, Technische Universität Chemnitz

Beuth, Frederik

Email: frederik.beuth@cs.tu-chemnitz.de

Dissertation: Visual attention in primates and for machines - neuronal mechanisms

Department of Computer Science, Technische Universität Chemnitz

Acknowledgments

I would like to express my deepest gratitude to all the people who helped me in the creation of this doctoral thesis. First and foremost, I would like to thank my graduate advisor and mentor, Prof. Fred Hamker, for sharing his great knowledge in the field of visual attention, his supervision, and continuous support of my work. Furthermore, I have to thank my second and third examiners, Prof. Heiko Neumann and Prof. Wolfgang Einhäuser-Treyer, for their interest in my work and their kindness to examine my thesis. This doctoral thesis would not be possible without the continued support from my family, colleagues, and friends over the years, thus I would like to express my sincere gratitude to them and thank them especially for their patience to bear with my long hours of work. Finally, I greatly have to thank the following persons for their invaluable help: Fanny Eichler and Julia Bergelt helped me extraordinarily to improve the graphical design of the figures, while Zampeta Kalogeropoulou, Marc Ritter, Arash Kermani, Julia Bergelt, Julian Vitay, and Hannah Leach were so kind to proofread my dissertation (or parts of it).

Abstract

Understanding visual attention is a long-term goal in the field of neuroscience, and a prerequisite to incorporate this essential concept of human perception into computer vision systems. However, understanding visual attention is challenging as it has many and seemingly-different aspects, both at neuronal (neurophysiological effects) and behavioral level (behavioral phenomena). Thus, it is very hard to give a uniform explanation of visual attention that accounts for all aspects. To tackle this problem, this doctoral thesis has the goal to identify a common set of neuronal mechanisms, which underlie the neurophysiological effects as well as the behavioral phenomena of attention. This approach allows it to explain both, neurophysiological and behavioral aspects with the same set of mechanisms. At first, the mechanisms are identified by modeling the neurophysiological effects of attention, and then these mechanisms are utilized to explain behavioral phenomena. In this way, also the mechanisms' roles in behavior are revealed. The mechanisms are simulated by neuro-computational models, thus leading to a neuro-computational framework that explains for the first time various aspects of attention at once. This allows for the neuroscience community to bridge the gap between neurophysiology and behavior, while the mechanisms provide for the computer vision community a concrete prototype implementation to incorporate attention into future systems. According to our work, attention would be very beneficial for computer systems as it provides for the brain a control process that tunes the visual system for the current task. This aspect of human perception is not present in current computer vision systems, thus an incorporation may certainly boost the current performances to new superior levels.

The first part of the thesis addresses the neurophysiology of visual attention. In each of the considered fields, we will also advance the current state-of-the-art to better understand this aspect of attention itself. The neurophysiological effects of attention are already well-replicated by computational models, but each model can typically account only for a few data sets. Thus, we develop here a novel microcircuit model of attention to explain a vast range of effects at once: spatial- and feature-based biased competition, modulation of the contrast response function, modulation of the neuronal tuning curve, and modulation of surround suppression. We identify four neuronal mechanisms as relevant: 1) amplification of neurons preferring an attended stimulus, 2) divisive normalization of neuronal responses, 3) spatial pooling within the receptive field of a neuron, and 4) suppression between neurons.

To model behavior, we scale up the microcircuit model to a novel system-level model, which simulates the visual cortex, the frontal eye field, and the prefrontal cortex. Thus, the model contains the same neuronal mechanisms as the microcircuit model, and simulates parts of the attentional processing network across the brain. Recent studies have indicated that this network controls neuronal activity for the current task in the whole visual cortex. This leads us to propose a more general view of visual attention: a cognitive and holistic control process, tuning the whole visual system for the task at hand.

As the first exemplary behavioral phenomenon, we simulate object localization, i.e. the task to search for a given target object in a scene, in the second part of the thesis. The task was chosen to illustrate for the computer vision community how the brain locates objects with the help of attention and to demonstrate the framework's robust applicability. Neuro-computational (or holistic) models avoid some drawbacks of the way attention is utilized until now in the computer vision community (saliency models), but have been only evaluated in small setups with a few objects and backgrounds, rising doubts about their applicability to realistic computer vision problems. Therefore, we show here at first that neuro-computational models are also applicable to realistic problems. For that, we demonstrate our model on a large task with 100 different target objects, 1000 scenes, and three background classes (black, white-noise, real-world). The high number of objects is possible as we introduce learned object descriptors. Our model achieves a localization accuracy of 92% at black backgrounds, at which the descriptors were learned. Generalization to white-noise or real-world backgrounds reduces the accuracies to 71% and 42% respectively. Secondly, we analyze the attentional processing in object localization at neuronal level. We know from neurophysiology that attention enhances (amplification mechanism) or suppresses neuronal responses (suppression), but not their roles in object localization. Putting the focus on feature-based attention, we found the following: The feature-based amplification mechanism represents the target in the visual system for the holistic control, while the feature-based suppression removes neuronal noise originating from other objects or the background, but is irrelevant for the performance in this task.

As a second behavioral phenomenon, we investigate object substitution masking (OSM) in the third part of the thesis to illustrate that the model can explain human behavioral data. OSM is a visual masking paradigm in which a target and a mask stimulus (four dots) are

briefly presented together. If the mask remains on the screen, the target cannot be recognized. Visual attention and OSM have been classically regarded as two separate phenomena, but OSM is here explained with our attention model, revealing that the phenomenon actually relies on attention mechanisms. Furthermore, the simulation of OSM and attention within a single model promotes the generality of our approach, and a neuro-computational model of OSM does not exist; hence, revealing the neuronal mechanisms of OSM is a huge advancement in itself. OSM is explained by our model through three attention mechanisms: The feature-based suppression mechanism accounts for the misrecognition of the target stimulus by inhibiting the target through the mask, the amplification mechanism explains how the target's neuronal representation is maintained after stimulus offset, and the dynamics of spatial attention accounts for the fact that OSM requires several distractors.

After the advancements in these subfields, we compare the role of the neuronal mechanisms in the individual attention aspects to link the various aspects of attention together. The neuronal mechanisms underly the neurophysiology of attention, and we found that they have the following roles in the behavioral phenomena: 1) The amplification mechanism has the behavioral role to represents the current task target in the visual system. 2) The suppression mechanism removes neuronal noise, or realizes a competition to select between several behavioral alternatives. 3) The mechanisms of normalization and spatial pooling play no relevant behavioral roles, at least in tasks investigated here.

To summarize, we explain many seemingly-different aspects of visual attention with a common set of neuronal mechanisms: amplification, divisive normalization, spatial pooling, and suppression. The three chosen phenomena, i.e. multiple neurophysiological effects, object localization, and object substitution masking, show that our approach can account for crucial neurophysiological, functional, and psychophysical properties of visual attention. As the elaborated neuronal mechanisms are sufficient to explain such a variety of aspects, we think they might constitute the essential neuronal substrate of visual attention in the cortex.

Kurzzusammenfassung

Visuelle Aufmerksamkeit ist ein essentielles Konzept der menschlichen Wahrnehmung. Es ist aber immer noch nicht komplett verstanden, so dass es ein langjähriges Ziel der Neurowissenschaften ist, das Phänomen grundlegend zu durchdringen. Gleichzeitig wird es aufgrund des mangelnden Verständnisses nur selten in maschinellen Sehsystemen in der Informatik eingesetzt, obwohl es essentiell für Menschen ist. Visuelle Aufmerksamkeit zu verstehen, ist allerdings eine komplexe Herausforderung, da das Phänomen viele verschiedene und verschiedenartige Aspekte hat, sowohl auf neuronaler Ebene (neurophysiologische Effekte) als auch auf der Ebene des menschlichen Verhaltens (Verhaltensphänomene). Daher ist es sehr schwierig, eine einheitliche Erklärung von visueller Aufmerksamkeit zu finden, welche für alle Aspekte gleichermaßen gilt. Um dieses Problem anzugehen, hat diese Arbeit das Ziel, gemeinsame neuronale Mechanismen von Aufmerksamkeit zu identifizieren, welche sowohl den neurophysiologischen Effekten als auch den Verhaltensphänomenen zugrunde liegen. Damit ist es mit einem Ansatz möglich, neurophysiologische und verhaltenstechnische Aspekte von Aufmerksamkeit auf einmal zu erklären. Zuerst werden dazu die neurophysiologischen Effekte von Aufmerksamkeit modelliert, um relevante neuronale Mechanismen heraus zuarbeiten, und dann werden diese Mechanismen benutzt, um Verhaltensphänomene zu erklären, sowie die Rolle dieser Mechanismen innerhalb selbiger zu untersuchen. Die Mechanismen werden in neuro-computational Modellen simuliert, was somit zu einem neuro-computational Modellierungsframework führt, welches zum ersten Mal verschiedenste Aufmerksamkeitsaspekte auf einmal erklären kann. Damit trägt das Framework zum Schließen der offenen Lücke in der Aufmerksamkeitsforschung zwischen Neurophysiologie und Verhalten bei. Für die Informatik stellt die Arbeit damit eine bessere und einheitlichere Beschreibung von visueller Aufmerksamkeit dar. Darüber hinaus liefert das Framework mit seinen neuronalen Mechanismen sogar eine Referenzimplementierung um Aufmerksamkeit in zukünftige Systeme integrieren zu können. Aufmerksamkeit könnte laut der vorliegenden Forschung sehr nützlich für diese sein, da es im Gehirn eine Aufgaben-spezifische Optimierung des visuellen Systems bereit stellt. Dieser Aspekt menschlicher Wahrnehmung fehlt meist in den aktuellen, starken Computervisionssystemen, so dass eine Integration in aktuelle Systeme deren Leistung sprunghaft erhöhen und eine neue Klasse definieren dürfte.

Der erste Teil der Arbeit modelliert die angesprochenen Auswirkungen von Aufmerksamkeit auf die neuronalen Feuerraten, untersucht somit die neurophysiologischen Effekte. Um Aufmerksamkeit auch in jedem Einzelaspekt besser zu verstehen, wird die Arbeit gleichfalls den aktuellen State-of-the-Art in jedem behandelten Bereich erweitern. Existierende Modelle können bereits zahlreiche neurophysiologische Effekte replizieren, allerdings kann ein einzelnes Modell typischerweise nur wenige Datensätze erklären. Darum wird in dieser Arbeit ein neues Microcircuitmodell mit dem Ziel entwickelt, sehr viele Effekte auf einmal erklären zu können. Dies umfasst: räumliche und merkmalsbezogene Biased-Competition, Modulation der Contrast-Response-Funktion, der neuronalen Tuning-Kurven, sowie von Surround-Suppression. Diese Modellierungsarbeit zeigt, dass die verschiedensten Effekte auf nur vier neuronalen Mechanismen beruhen, welche damit quasi die für Aufmerksamkeit relevante neuronale Verarbeitung darstellen: 1) Erhöhung der Feuerrate (Amplifikation) von mit Aufmerksamkeit assoziierten Neuronen, 2) Divisive Normalisierung der neuronalen Feuerrate, 3) Pooling über einen räumlich definierten Bereich, und 4) Dämpfung (Suppression) von neuronalen Reaktionen.

Um Verhaltensphänomene zu modellieren, wird das Microcircuitmodell anschließend auf ein Systemmodell von visueller Aufmerksamkeit hochskaliert, welches den visuellen Cortex, das Frontale Augenfeld und den präfrontalen Cortex enthält. Damit enthält das Systemmodell die gleichen neuronalen Mechanismen wie das Microcircuitmodell, und simuliert Teile des Aufmerksamkeitsnetzwerkes des Gehirns. Neuere neurowissenschaftliche Forschungsergebnisse zeigen, dass dieses Netzwerk die neuronale Aktivität Aufgabenspezifisch und im gesamten visuellen Cortex kontrolliert. Aus diesen Gründen wird hier vorgeschlagen, Aufmerksamkeit als einen kognitiven und holistischen Kontrollprozess zu sehen, der das gesamte visuelle System für die aktuelle Aufgabe justiert.

Im zweiten Teil der Arbeit wird als erstes Verhaltensphänomen Objektlokalisation simuliert, daher die Aufgabe ein vorgegebenes Objekt in einer Szene zu finden (auch als Guided-Visual-Search bekannt). Das Phänomen wurde ausgewählt, um für die Computervision Community zu demonstrieren, wie das menschliche Gehirn Objekte lokalisiert, und um die robuste Funktionalität des Frameworks aufzuzeigen. In der Computervision Community wurde Aufmerksamkeit bislang meist nur als räumliche Vorselektion eingesetzt (Salienzmodell), was einige Nachteile mit sich bringt. Diese werden von neurocomputationalen Modellen vermieden, allerdings wurden diese Modelle trotz ihrer Vorteile bislang nur in kleinen Aufgaben mit wenigen Objekten oder Hintergründen eingesetzt. Deswegen illustriert die Arbeit hier als erstes, dass neuro-computationale Modelle auch für reale Computervision-Anwendungen einsetzbar sind. Dazu wird das Modell an einem großen Setup mit 100 Objekten, 1000 Szenen und 3 verschiedenen Hintergrundklassen (Schwarz, Noise, Real-Welt) demonstriert. Die hohe Anzahl an Objekten ist möglich, da als Neuheit lernbare Objektdeskriptoren eingesetzt werden. Die Objektdeskriptoren wurden auf den schwarzen Hintergründen gelernt und das Modell erreicht auf diesem nativen Setup eine Lokalisationsrate von 92%. Falls es auf die anderen Hintergründe generalisieren muss, reduziert sich die Rate auf 72% (Noise) bzw. 42% (Real-Welt). Anschließend wird die Aufmerksamkeitsverarbeitung auf neuronaler Ebene untersucht. Aus der Neurophysiologie ist bereits bekannt, dass Aufmerksamkeit die neuronalen Feuerraten erhöht oder senkt (Amplifikations- bzw. Suppressionsmechanismus). Da aber bislang nicht bekannt ist, welche Rolle ein einzelner dieser Mechanismen bei der Objektlokalisation spielt, erfolgt eine Untersuchung desselben (mit einem Fokus auf merkmalsbezogene Aufmerksamkeit). Es resultieren die Erkenntnisse: Der (merkmalsbezogene) Amplifikationsmechanismus enkodiert das Zielobjekt im visuellen System, so dass Aufmerksamkeit das visuelle System für die aktuelle Aufgabe justieren kann. Der (merkmalsbezogene) Suppressionsmechanismus unterdrückt neuronales Rauschen, stammend z.B. vom Hintergrund oder anderen Objekten, aber er ist irrelevant für die Performance in dieser Aufgabe.

Im dritten Teil der Arbeit wird als zweites Verhaltensphänomen Object-Substitution-Masking (OSM) untersucht, um zu zeigen, dass das Framework menschliche Verhaltensdaten replizieren kann. OSM ist ein Paradigma, in welchem die Erkennbarkeit eines kurz gezeigten Zielstimulus durch einen parallel und länger gezeigten Maskenstimulus reduziert wird (Maskierungsparadigma). OSM und visuelle Aufmerksamkeit wurden bisher als zwei getrennte Wahrnehmungsphänomene gesehen. Allerdings wird OSM hier mit einem Aufmerksamkeitsmodell erklärt, und zeigt damit auf, dass dies – anders als bisher angenommen - auf Aufmerksamkeitsmechanismen beruht. Darüber hinaus stellt die Simulation von OSM und Aufmerksamkeit in einem Modell nochmals die breite Allgemeingültigkeit des vorgeschlagenen Ansatzes heraus. Desweiteren existiert im Moment kein neurocomputational Modell für OSM, so dass die Enthüllung der neuronalen Verarbeitung in OSM schon ein großer Fortschritt für sich selbst ist. Die vorliegende Modellierungsstudie zeigt, dass OSM primär auf drei neuronalen Aufmerksamkeitsmechanismen beruht: Der merkmalsbezogene Suppressionsmechanismus erklärt die Misserkennung des Zielstimulus durch die Unterdrückung des Ziels durch die Maske. Der Amplifikationsmechanismus ermöglicht die interne Weiterrepräsentation des Zieles im Gehirn nach Abschalten des Stimulus. Die Dynamik von räumlicher Aufmerksamkeit erklärt letztlich, warum OSM mehrere zusätzliche Objekte auf dem Bildschirm (Distraktoren) benötigt.

Abschließend wird die Rolle der neuronalen Mechanismen in den einzelnen Aufmerksamkeitsaspekten verglichen, um die verschiedenen Aspekte von Aufmerksamkeit miteinander zu verknüpfen. Zu diesem Zweck wird die Rolle der neuronalen Mechanismen, welche ja aus der Neurophysiologie stammen, in den Verhaltensphänomenen erfasst. Es ergeben sich folgenden Rollen in den Verhaltensphänomenen: 1) Der Amplifikationsmechanismus enkodiert das aktuelle Aufgabenziel im visuellen System. 2) Die Suppression hat den Zweck neuronales Rauschen zu entfernen, oder einen Wettbewerb zwischen Verhaltensalternativen zu implementieren. 3) Die Normalisierung und das Pooling haben keine hervorstechende Rolle auf Verhaltensebene, zumindest in den zwei hier untersuchten Experimenten.

Zusammengefasst kann die Arbeit viele scheinbar verschiedene Aspekte von visueller Aufmerksamkeit mit einem gemeinsamen Set von neuronalen Mechanismen erklären: Amplifikation, Divisive Normalisierung, räumliches Pooling, und Suppression. Die drei gewählten Phänomene, d.h. Neurophysiologie, Objektlokalisation, und Object-Substitution-Masking, zeigen, dass der Ansatz grundlegende neurophysiologische, funktionale und verhaltenstechnische Eigenschaften von visueller Aufmerksamkeit erklären kann. Da die gefundenen Mechanismen somit ausreichend sind, Aufmerksamkeit so umfassend erklären zu können, könnten sie vielleicht sogar das essentielle neuronale Substrat von visueller Aufmerksamkeit im Cortex darstellen.

Preface

The author has published several studies that contributed to this work, or are part of it:

The first part of this thesis (Chapter 3: “Microcircuit model of attention”) has been published in full as:

- Beuth, F. and Hamker, F. H. (2015a). A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Res*, 116(Part B):241–57.

The second part (Chapter 4: “Object localization with a model of visual attention”) has been published in parts as:

- Beuth, F, Hamker, F. H. (2015b). Attention as cognitive, holistic control of the visual system. In *Proc Workshop New Challenges in Neural Computation 2015 - NCNC 2015*, pages 133-140.
- Beuth, F, Hamker, F. H. (2015). Object localization with a neurophysiologically-precise model of visual attention. *Invited presentation at the Symposium on Attention and Cognitive Systems - ISACS 2015, associated to the IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS 2015*.
- Beuth, F, Hamker, F. H. (2015). Object recognition and visual search with a physiologically grounded model of visual attention. *Presentation at the Workshop for Computational and Mathematical Models of Vision - MODVIS 2015*,

and is based on the following earlier studies, containing system-level models of attention, of the author:

- 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.
- Beuth, F., Jamalian, A., and Hamker, F. H. (2014). How Visual Attention and Suppression Facilitate Object Recognition? In *Proc 24th Int Conf Artificial Neural Networks - ICANN 2014*, pages 459–466.

- Zirnsak, M., Beuth, F., and Hamker, F. H. (2011a). Split of spatial attention as predicted by a systems-level model of visual attention. *Eur J Neurosci*, 33(11):2035–45.
- Beuth, F., Wiltschut, J., and Hamker, F. H. (2010). Attentive stereoscopic object recognition. In *Proc Workshop New Challenges in Neural Computation 2010 - NCNC 2010*, pages 41–48.

The third part (Chapter 5: “Object substitution masking”) has been presented as posters at several international conferences:

- 15th Annual Meeting of the Vision Sciences Society - VSS 2015.
- 37th European Conference on Visual Perception - ECVP 2014.
- International Workshop ’Neuro-Cognitive Mechanisms of Conscious and Unconscious Visual Perception’ - CUVP 2014.
- 36th European Conference on Visual Perception - ECVP 2013.

Contents

List of Figures

List of Tables

List of Abbreviations

COIL-100	Columbia Object Image Library - 100 object database
cRF	Classical receptive field
CRF	Contrast response function
FEF	Frontal eye field
FEFv	Visual cells in the FEF
FEFvm	Visuomovement cells in the FEF
FEFm	Movement cells in the FEF
FEFfix	Fixation cells in the FEF
HVA	Higher visual area
IT	Inferior temporal cortex
LGN	Lateral geniculate nucleus
LMS	Colorspace encoding long(L), middle(M) and short(S) wavelengths
LIP	Lateral intraparietal area
MST	Medial superior temporal cortex
MT	Medial temporal cortex
7a	Brodmann area 7a
OSM	Object substitution masking
PFC	Prefrontal cortex
RF	Receptive field
RT	Reaction time
V1	Primary visual cortex
V2	Secondary visual cortex
V3	Third visual cortex
V4	Fourth visual cortex

1. General introduction

1.1. Motivation

Visual attention is an important cognitive concept for the daily life of humans, but it is still not fully understood. The concept describes the human ability to focus oneself to a particular aspect of a scene, like an object or a location. This is very useful in daily life, as for many tasks, only a few objects are really important. For example, when humans search for a particular object on an overstuffed office desk, only the target object is relevant while all other objects on the desk can be safely ignored (these are often called distractors). Attention is directed to the target object in this task, which focuses cognitive resources on the processing of the target object, and filters out from the incoming sensory data the task-irrelevant distractor objects. Those aspects led to common definitions of visual attention: Hamker (2005b) defines it as the cognitive process to focus neuronal processing resources on an aspect of a scene, while Carrasco (2011) sees it as the selection of relevant information among the vast amount of sensory data.

In this doctoral thesis, we aim at a deeper understanding of visual attention by elaborating its underlying neuronal mechanisms. Visual attention has been studied for decades in multiple disciplines, revealing that it has many numerous aspects. For example, a recent review of visual attention in 2011 (Carrasco, 2011) pointed out that a literature search alone in PubMed yields up to 2400 articles, with the tendency of an exponential increase. This huge amount of findings makes it currently very hard to give a uniform explanation of attention. Models and theories are currently not able to keep up with the enormous findings, and are typically only able to account for a single or a few data sets. This motivates us to develop models that are capable to explain more uniformly the different aspects of attention. Our idea to tackle this problem is here to work out common neuronal mechanisms that are able to explain various data sets. As many findings stem from studying the same anatomical structure, the visual cortex, we think common mechanisms should exist. The literature shows that attention modulates both neuronal responses *(neurophysiological effects, Reynolds and Heeger (2009))* as well as human behavior *(behavioral phenomena, Carrasco (2011))*. To simulate both in our work, we firstly identify the neuronal mechanisms based on the neurophysiological effects of attention. Following this, we then reveal the role of these mechanisms in behavioral phenomena, and in such a way, explain a particular phenomenon through these mechanisms. This results in a neuro-computational framework that advances in each field the state-of-the-art, but also utilizes the same mechanisms, and thus ensures a uniform explanation of the different attention aspects.

This research course has besides a more uniform grasp of visual attention the immediate advantage that it bridges the gap between the neurophysiological and behavioral aspects of attention. The link between these two remained at moment rather vague, for example the role of a particular neurophysiological effect at behavioral level is still mostly unclear. Bridging both sides is complicated because neurophysiological and behavioral experiments use different setups, and the experiments measure different outcomes (neuronal responses versus behavioral decisions). Our approach for bridging this gap is to trace both neurophysiological and behavioral aspects back to the same underlying neuronal mechanisms. This course of action does not link a particular neurophysiological effect to a particular behavioral phenomenon as this is hardly possible, instead it links both aspects to a common set of underlying neuronal mechanisms. At the end of this thesis, we then compare which mechanism is involved in which effect and which phenomenon, and in this way, infer indirectly the link from neurophysiology to behavior.

Besides this neuroscientific incitement, our work is equally motivated from the perspective of computer science. In our view, visual attention is a potential superior concept for future computer vision systems. Attention is involved in almost any visual perception phenomena (Carrasco, 2011) and thus it seems to be an essential part of the human vision system. Especially interesting for computer vision systems is that attention modulates neuronal activities (Reynolds and Heeger, 2009), and that this modulation is task-specific as attention is per definition task-specific. Due to these and other findings, we see attention as a cognitive control process, tuning the visual system for the current task. For example in a recognition task, this task-dependent tuning may allow to filter out task-irrelevant objects (distractors) directly within the visual system by inhibiting their representations, which might otherwise disturb the recognition process. This may be a superior approach to current computer vision systems that recognize all objects including the distractors in a visual stage, and subsequently filter out the distractors. However, visual attention has been rarely utilized in computer vision systems despite its benefits. One reason may be that it is currently unclear how to implement attention correctly because a uniform and deep understanding of visual attention up to the neuronal level is not available. Our work will provide a deeper understanding, and even more, a set of neuronal mechanisms that can serve as a prototype to incorporate attention into future computer vision systems. Therefore, a powerful future computer vision system might be based on two pillars: a sophisticated representation of visual stimuli and a fine-grained attentive control. Powerful stimulus representations have been already learned in recent years by deep neuronal networks (Zeiler and Fergus, 2014). Attention is much less understood, hence we see this topic as a very promising research direction.

Our thesis will start with modeling the neurophysiological effects of attention. One of the first discovered effects was biased competition (Desimone and Duncan, 1995): If two different stimuli are presented closely together, they will suppress each other, which results in a moderate neuronal response. Directing attention to one of these stimuli will increase the response of neurons preferring this stimulus *(attended stimulus)*, and will decrease the response of neurons preferring the other stimulus without attention *(unattended stimulus)*. Besides biased competition, attention has many other neurophysiological effects. For example, it affects in multiple ways the tuning of neurons for their preferred stimulus, it also increases/decreases the response of neurons when suppressed from stimuli in the far surround, and its response modulation depends in several ways on the strength of the visual input (attentional modulation of the neuronal tuning curve, surround suppression, and contrast response function, respectively). Existing models have already replicated many findings, but each model has typically accounted only for a few of them. Our contribution in this field is that we develop a novel model that can explain multiple data sets at once by utilizing a common set of neuronal mechanisms. The mechanisms are to some degree already present in current models, yet we unify the mechanisms, add some new ones and underpin them with neuroanatomical details as most existing models also lack a neuroanatomical foundation. Our model simulates a single cortical area, thus we called it single-area or microcircuit model. With the novel model, we are able to explain many of the existing effects (12 different data sets) at once. We will find that four neuronal mechanisms are relevant for attention: amplification, spatial pooling, divisive normalization, and suppression. *Amplification* simulates the response amplification of neurons preferring the attended stimulus. *Spatial pooling* describes a pooling within a local area of a neuron. This operation is typically found in the visual cortex (Hamker and Zirnsak, 2006), and we will argue that it is also necessary for attention, especially for biased competition. The next mechanism, the *divisive normalization* of neuronal responses (Reynolds and Heeger, 2009), accounts for the contrast response function and its attentional modulation. Lastly, the *suppression* mechanism models the observed response decrease of neurons, for example of neurons preferring the unattended stimulus in biased competition or surround suppression.

Afterwards, we will investigate the role of the neuronal attention mechanisms in behavioral phenomena. For this, we scale up the microcircuit model to a system-level model, containing major relevant areas for attentional processing: a lower and a higher visual area, the prefrontal cortex, and the frontal eye field. Some recent studies have shed light on the attentional processing at system-level. Miller and Buschman (2013) have illustrated that attentional processing spans a top-down network from prefrontal cortex to visual cortices as well as to spatial cortices like the frontal eye field. The prefrontal cortex is involved in task control (Sakai, 2008), thus this network may transport task-relevant information to the visual system. From neurophysiology, we know that attention amplifies neurons preferring the attended stimulus, and suppresses others. The attended stimulus is per definition task-relevant, thus we reason that attention amplifies the response of neurons preferring task-relevant stimuli, and suppresses other neurons. Putting these evidences together leads us to propose a novel view of attention: a cognitive and holistic control process, tuning the visual system for the current task. We termed it as holistic to emphasize that the process modulates activity for a single purpose, the task, in the whole visual cortex in parallel.

As the first behavioral phenomenon, we investigate object localization due to its high relevance for computer vision and robotics. We analyze here the object localization task to localize a given object in a real-world scene, known as attention phenomenon of guided visual search (Wolfe, 1994). Current computer vision approaches cannot solve this problem perfectly, so we would like to propose the holistic attention approach of the primate brain as alternative. The computer vision community also uses already attention to solve this task, namely in the saliency models (Borji and Itti, 2013). Yet, such models utilize attention solely as a spatial pre-selection stage for a subsequent recognition module. This trivial approach has two issues that we would like to avoid. Firstly, attention and recognition operate in separate stages, but should be intertwined (Frintrop et al., 2010) as attentional selection and recognition depend on each other. Our holistic approach controls selection and recognition in parallel, solving the interdependency problem. Secondly, the attentional pre-selection stage operates only correctly if the target is conspicuous in the scene (bottom-up saliency models) or distinguishable by some simple feature like color (top-down saliency models). The holistic approach avoids such limitations as it distinguishes the target via high-level object representations rather than via simple features. However, the existing neuro-computational models, or in our terms holistic models, have been rarely applied to real-world object localization tasks, i.e. to tasks with whole objects and real-world scenes, which are common in the computer vision domain. The reason behind this is that many models have been developed to investigate guided visual search, which uses typically very simple setups like red and green bars before a black background. Thus, merely a few models have been applied to real-world objects localization tasks yet: Antonelli et al. (2014), Beuth et al. (2010), Chikkerur et al. (2010), Hamker (2005b), and Walther and Koch (2007). Two of the models, Antonelli et al. (2014) and Beuth et al. (2010), were developed by the author himself during an earlier project.

Our contribution in this field is that we develop the holistic object localization approach further, especially regarding its applicability to computer vision tasks. At first, we improve shortcomings of the existing models. Our model stems primarily from the model of Hamker (2005b) as it has a much better neurophysiological and neuroanatomical grounding than the other real-world capable models. However, it has also disadvantages that we would like to improve: It cannot learn object representations, thus we introduce learned representations which allows to encode numerous objects as required for computer vision tasks; its frontal eye field is slightly dated, thus we update it; and of course we use our neuronal mechanisms which allows to replicate more neurophysiological attention effects than shown previously (Hamker, 2005a). Secondly, we demonstrate with this novel model that neuro-computational (or holistic) models are applicable to realistic computer vision problems. Only a few current models have been applied to real-world problems yet, and to matters worse, these applications have used only a low number of object categories and backgrounds. This raises massive doubts about the applicability of neuro-computational models to realistic computer vision problems. Therefore, we demonstrate our model on such a problem, namely a large and realistic task with 100 objects, 1000 scenes, and three background classes. Finally, we illustrate the attentional processing in object localization and analyze the roles of our neuronal attention mechanisms in it. The existing models have already proposed explanation for the attentional processing in object localization. Here, we will follow and improve the one of Hamker (2005b). We will at first illustrate again the attentional processing in object localization as this seems not widely-known in the community. Next, we will advance the existing explanations especially regarding the individual role of each neuronal mechanism in the processing. It is already known from neurophysiology that attention can enhance or suppress neurons (amplification or suppression mechanisms), but the individual role of such a mechanism in object localization has remained vague as the existing models invoke typically multiple mechanisms in parallel, leaving the role of a single mechanisms unclear. As such knowledge describes the implementation level of attention in the brain, it would be very desirable to know for a future incorporation of attention into computer vision systems.

As second phenomenon, we investigate object substitution masking (OSM) to illustrate that our modeling framework can explain psychophysical data. Psychophysics is a subfield of psychology which studies the human perception of physical stimuli (Gescheider, 1997), thus we would like to demonstrate that our modeling approach can explain human perceptional and behavioral data. OSM is one of the psychophysical paradigms, belonging to the group of visual masking paradigms, which characterize experiments to reduce the visibility of a target stimulus with a mask stimulus, and are mostly used to study the temporal aspects of visual perception (Breitmeyer and Ogmen, 2006). In OSM, the target stimulus is a ring with a gap, which is enclosed by four dots serving as the mask stimulus. At first, both stimuli are presented together for a very brief period of time, and then either the mask is set off with the target or it remains on the screen. The target is well recognizable in the first condition despite its short presentation period, but not in the second when the mask stays on the screen. In this latter condition, the experiment's participants reported that the mask seems to replace (or substitute) the target in their perception, leading to the name of the phenomenon.

In this part of the thesis, we explain OSM fully with our attention model, showing that OSM actually relies on the same neuronal mechanisms than visual attention. This bridges the gap between the two seemingly-different phenomena of OSM and visual attention, which are classically thought to rely on different mechanisms, despite discussions about OSM being affected by attention. As it is also still under debate which of the existing OSM theories truly accounts for the phenomenon (Goodhew et al., 2013), we herewith would like to propose visual attention as an alternative explanation for OSM. Besides this contribution, we choose OSM also as currently no computational-neuroscience model of it exists, thus revealing neuronal mechanisms in OSM is a huge advancement in itself. Finally, a masking experiment was chosen as it requires realistic temporal processing in the model, allowing us to expand the model in the future to many other psychophysical paradigms and deepen our understanding of the temporal communication within the visual cortex. Artificial computer vision systems would also benefit from this understanding, as their communication could be improved by omitting unnecessary messages. This would enable us to speedup existing systems or to create larger ones.

In the last part of the thesis, we analyze the role of the neuronal mechanisms in each field to compare the attentional processing across the neurophysiological effects and behavioral phenomena. Interestingly, the two phenomena constitute fundamentally different tasks, namely a localization and recognition task. This allows us to investigate attentional processing in two fundamentally different tasks classes and to work out the differences. We define both task classes as follows (similarly to Tsotsos et al. (2005)): In a *localization task*, the type of the target stimulus is given and its location has to be determined. In contrast in a *recognition task*, the location of the target is given and its type has to be determined. OSM is a recognition task as the location of the target stimulus is given by the four-dot mask and the target's type has to be determined during the recognition. Our investigation of both tasks allows now to compare the roles of our neuronal attention mechanisms in both tasks and to predict the relevance of the neuronal mechanisms for different task classes. We will precisely conduct this in the last part of the thesis.

To summarize, this doctoral thesis will work out common neuronal mechanisms of attention to explain many seemingly-different aspects of visual attention at once. For each aspect, we will elaborate how the mechanisms operate and also advance the state-of-the-art in this field. The chosen aspects, i.e. multiple neurophysiological effects, real-world object localization, and OSM, highlight that our mechanisms can account for crucial neurophysiological, functional, and psychophysical properties of visual attention. As the mechanisms account for attention under such a variation of view points, they might constitute the general neuronal substrate of attention in the visual cortex.

1.2. Outline

To start, we give the reader in Chapter 2 an overview of visual attention, illustrating the manifold different aspects of the phenomenon and the capabilities of the current models to account for them. It will show that models are already good at replicating single aspects, but fail to account for multiple ones. Afterwards, this thesis is structured from small to large scale attentional processing. In Chapter 3, attentional processing is investigated at neurophysiological level, showing how the microcircuit model of attention explains a very large set of single-cell recordings. The chapter in full has been published as Beuth and Hamker (2015a). Based on this small model and its mechanisms, we build a system-level model by including major cortical areas involved in attentional processing. This system-level model is used in Chapter 4 to perform object localization with the help of attention, and in Chapter 5 to explain OSM with attentional processing. Therefore, the Chapters 3, 4, and 5 illustrate that different visual attention aspects can be explained by one unified model, which is only possible because the processing relies on common attention mechanisms. We conclude this in chapter 6 and summarize which mechanism is involved in which phenomenon.

2. The state-of-the-art in modeling visual attention

The phenomenon of visual attention has been investigated for several decades and in a very broad range of studies. Resulting modeling work is also wide and distributed over several disciplines. In the following, we give an overview of the models of visual attention. To make the huge amount of modeling literature accessible for the reader, we approach the field from three perspectives (Fig. 2.1): neurophysiology, applied systems for object localization and recognition, and system-levels models for psychophysics. There exist some high-quality reviews in each field, which we will use as a starting point for our own summarization in the following. The reviews have appropriate these foci: Neurophysiological findings have been reviewed by Reynolds and Heeger (2009); computer vision systems for object localization (sometimes also called recognition) by Frintrop et al. (2010), Borji and Itti (2013), Filipe and Alexandre (2013), and Jamalian and Hamker (2016); psychophysical studies by Carrasco (2011); and models of all three perspectives by Tsotsos and Rothenstein (2011).

We will focus in this thesis on computational models. One particular advantage of them is their ability to reveal underlying mechanisms and so profoundly explain a phenomenon. Hence, this chapter will focus strongly on computational models, but of course will also explain the empirical data on the way along. The interest in visual attention has strongly increased over the last 15 years (Carrasco, 2011). A lot of new discoveries were made, especially in the area of neurophysiology. Also, a lot of new models were developed in the field, following the discoveries, or finer elaborating older data. Due to these growing interests and a lot of new studies, we focus in this work especially on the recent and latest developments in the field (since about the year 2000). On the other hand, the research on visual attention has a very long tradition. The author is aware that some of the greatest psychologists have worked in the field, like William James, Wilhelm Wundt, and Anne Treisman (also Donald E. Broadbent, Michael Posner, Ulric Neisser). A deep historical development can be found in the background of visual attention literature that have tried to give definitions for the mechanisms of attention from different psychophysical, cognitive,

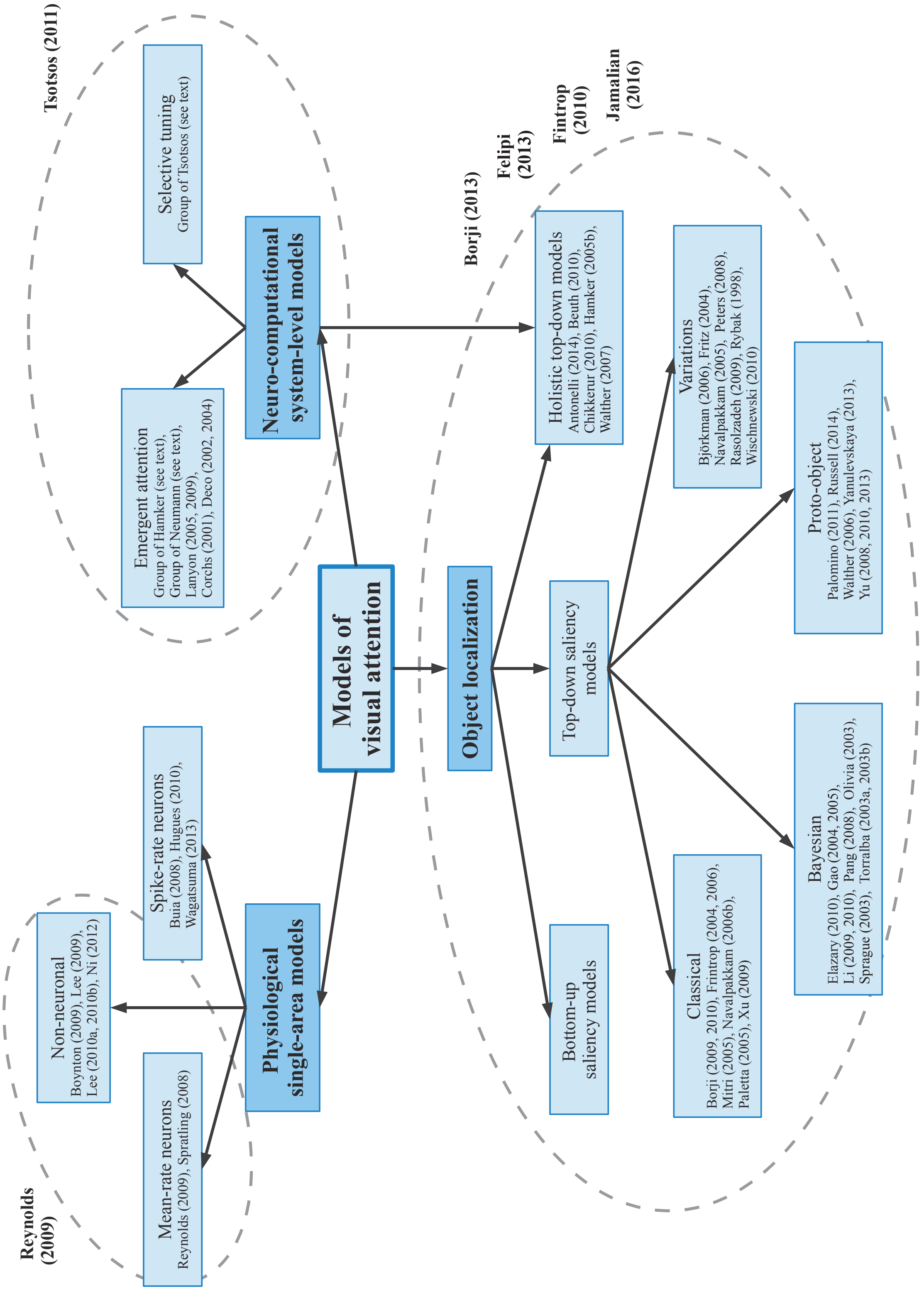

Models of visual attention
Neuro-computational system-level models
Tsotsos (2011)
Selective tuning
Group of Tsotsos (see text)
Emergent attention
Group of Hamker (see text),
Group of Neumann (see text),
Lanyon (2005, 2009),
Corchs (2001), Deco (2002, 2004)
Physiological single-area models
Reynolds (2009)
Spike-rate neurons
Buia (2008), Hugues (2010),
Wagatsuma (2013)
Non-neuronal
Boynton (2009), Lee (2009),
Lee (2010a, 2010b), Ni (2012)
Mean-rate neurons
Reynolds (2009), Spratling (2008)
Object localization
Borji (2013)
Felipi (2013)
Fintrop (2010)
Jamalian (2016)
Holistic top-down models
Antonelli (2014), Beuth (2010),
Chikkerur (2010), Hamker (2005b),
Walther (2007)
Top-down saliency models
Bottom-up saliency models
Variations
Björkman (2006), Fritz (2004),
Navalpakkam (2005), Peters (2008),
Rasolzadeh (2009), Rybak (1998),
Wischnewski (2010)
Proto-object
Palomino (2011), Russell (2014),
Walther (2006), Yanulevskaya (2013),
Yu (2008, 2010, 2013)
Bayesian
Elazary (2010), Gao (2004, 2005),
Li (2009, 2010), Pang (2008), Olivia (2003),
Sprague (2003), Torralba (2003a, 2003b)
Classical
Borji (2009, 2010), Frintrop (2004, 2006),
Mitri (2005), Navalpakkam (2006b),
Paletta (2005), Xu (2009)

Figure 2.1.: A mind map of recent visual attention models. The ellipses denote reviews. All publications are cited here by the first author, while the full reference is given in the main text.

and neurophysiological aspects. However, here we have presented these descriptions from the aspects of neuro-computational modeling work, and for a more general review of these historical bases, we would refer the reader to one of the very good reviews in the community. For example, Carrasco (2011) gives an excellent and recent overview of visual attention, which we will use to outline the phenomenon of attention before starting the modeling work. If of interest, we also give the list of historical main works for the reader: Von Helmholtz (1867, p. 741), James (1890, Chap. 11), Wundt (1874, Chap. 18), Treisman (1964) and Treisman and Gelade (1980); (Broadbent (1958), Posner (1980), Neisser (1967)); especially, as some precise references are not listed by newer reviews like Carrasco (2011). Regarding models, a review of earlier models can be found in the following works: Hamker (1999), Frintrop et al. (2010), Logan (2004). The first two provide general reviews focusing on computational models, while Logan (2004) focuses on formal, psychophysical theories of visual attention, in time from the 1950s up to today.

Attention can be modeled at the level of a single cortical area (*single-area model*) or at the level of multiple areas (*system-level model*). The first variant is typically utilized to explain neurophysiological findings, whereas the latter primarily explains psychophysical findings, rarely neurophysiological data, and is sometimes used for object localization.

In this chapter, we will first give a general introduction of the phenomenon of visual attention (Sec. 2.1), and clarify the involved brain structures (Sec. 2.2). Afterwards, we will explain visual attention via the three perspectives and give the reader an overview of computational modeling approaches in each: single-area models to explain neurophysiological data (Sec. 2.3), visual attention in object localization and recognition systems (Sec. 2.4), and neuro-computational system-level models to explain psychophysics or system-wide attentional processing (Sec. 2.5). Each of the perspectives will form the foundation for one chapter of this doctoral thesis (Chap. 3 - 5).

2.1. The phenomenon of visual attention

In the following, we will characterize briefly the phenomenon of visual attention, introduce common terms, and differentiate it from attention outside the visual domain. Visual attention can be defined as the concept to focus cognitive resources on an aspect of a scene (Hamker, 2005b), or as the selection of relevant information among the vast amount of sensory data (Carrasco, 2011). It can be deployed either to a feature *(feature-based attention)* or to a location *(spatial attention)*, whereby the experimental paradigm or task typically determines which kind of attention is performed. Feature-based attention is defined as attending to a certain feature, such as a specific edge, object or motion, whereas spatial attention implies to attend to a certain location. Feature-based attention includes in our understanding also attending to a certain object which is by some authors seen as a separate type of attention (object-based attention, Roelfsema et al. (1998)), because an object can be seen as a very complex feature.

Visual attention can further operate without eye movements (*covert attention*, Carrasco (2011)), or together with them, in which case the eye movements are typically performed to the attended location *(overt attention)*. Visual attention can be further classified into top-down and bottom-up visual attention (Carrasco, 2011; Ward, 2008). *Top-down visual attention* describes the phenomena that attention emerges top-down from our mind, driven by the task's demands and goals. It can be deployed voluntary dependent on the current task goals, and selects the relevant sensory information for the current task. Alternative terms are selective visual attention, nowadays also used in the field of neurophysiology (Busse, 2006), and endogenous attention or sustained attention used mainly in psychophysics (Carrasco, 2011; Ward, 2008). Next to top-down attention, the literature knows a stimulus-driven form of attention, denoted as *bottom-up visual attention*. It describes the phenomenon that animals or humans orient themselves automatically to a suddenly-appearing or otherwise conspicuous stimulus, which goes along with a deployment of attention. This form is also called exogenous or transient attention. Bottom-up attention is investigated more often in the field of the psychophysics, whereas neurophysiologists have examined habitually more the goal-driven form of visual attention, i.e. top-down attention (Reynolds and Heeger, 2009). This thesis primarily addresses the topic of top-down visual attention, thus this literature overview is more focused on this kind. Due to this focus, the term "visual attention" refers in this thesis always, if not otherwise noted, to top-down visual attention, and is also shortened to that term for the sake of brevity. Top-down attention is also simulated by the majority of models (Sec. 2.3, 2.4.2, 2.4.3, 2.5), often together with bottom-up processing, and only a single branch of the models in object localization uses solely bottom-up attention (Sec. 2.4.1, bottom-up saliency models).

Visual attention might be part of a larger network of attention systems in the human brain as attention is known also from other domains. For example, attention is known from other sensory domains like audio, where it operates similarly as a selection process (auditory attention, Fritz et al. (2007)). It is also known from emotional processing, where nuclei involved in emotional processing provide the emotional relevance of a stimulus, and attention is then guided to such stimuli (emotional attention, Liebold et al. (2015)). Lastly, attention is also investigated in the context of behavior in general, which is denoted as executive attention. A very influential theory stems from Posner and Peterson (Posner and Petersen, 1990; Posner and Rothbart, 2007; Petersen and Posner, 2012), which proposes that the brain uses three attention systems: an alerting, orienting, and executive system (terms according to Petersen and Posner (2012)). The alerting system is responsible for maintaining a behavioral state of a high sensitivity for incoming stimuli, the orienting system for selecting task-relevant information from the sensory input, and the executive system for the execution of behavior which includes the resolving of conflicts or performing of task switches as well. Visual attention might be a part of these other attention systems, but the precise relationship between them is not known yet. However, as other forms of attention may certainly have an influence on visual attention, studies denote sometimes phenomena from these others forms as visual attention phenomena. A typical example may be when emotional attention is investigated in the visual domain. In this doctoral thesis, we will leave out such mixed forms, as our thesis solely aims at the modeling of visual attention.

2.2. Brain structures involved in visual attention

Visual attention phenomena are prominently observed in all areas of the visual cortex and in a few areas of the frontal cortex, e.g. FEF and PFC. In the following, we will briefly characterize the relevant areas and the processing.

The visual cortex is comprehensively reviewed in several works (Chinellato, 2008; Felleman and Van Essen, 1991; Jamalian and Hamker, 2016; Orban, 2008; Serre, 2006; Ungerleider and Haxby, 1994). We would like to refer the reader to one of these reviews, e.g. Orban (2008), if he or she is unfamiliar with the visual cortex as we will recapitulate this topic here only very briefly. In this work, we will use the area nomenclature of the human cortex (Jamalian and Hamker, 2016), which differs slightly from other primate species like the macaque monkey (Felleman and Van Essen, 1991).

The visual attention system of humans is depicted as a schematic overview in Fig. 2.2a, showing the system's key connections and functions. We limit ourselves to them as the actual wiring would be much too complicated. For example, Felleman and Van Essen (1991) illustrate the actual wiring for the macaque monkey's visual system (Fig. 2.2b).

The processing in the visual attention system consists of a bottom-up processing of visual information and of a top-down control by attention, originating from high-level cortices. The processing of visual information starts with the retina (Wässle and Boycott, 1991) which perceives the visual world via photoreceptors at multiple wavelengths (Bowmaker and Dartnall, 1980), and converts the visual information into neuronal responses (retinal ganglion cells, Gegenfurtner (2003); Chatterjee and Callaway (2003)). These responses are transferred to the lateral geniculate nucleus (LGN, Wiesel and Hubel (1966)) in the thalamus and afterwards to the visual cortex, consisting prominently of the areas V1, V2, V3, V4, IT, MT, MST, 7a, and LIP. The visual cortex is organized according to Ungerleider and Haxby (1994) in two processing streams, characterized as "what"- and "where"-streams. The ventral stream processes the type of an object ("what", green in Fig. 2.2a) and uses the areas V1, V2, V4, and IT, while the dorsal stream processes the location of an object ("where", blue) and invokes the areas V1, V2, V3, MT, MST, and 7a. A high-quality review about the ventral stream can be found in Serre (2006), about the dorsal stream in Chinellato (2008). Neural information are also exchanged between the streams (Gazzaniga et al., 2009), despite they are viewed as separated units by Ungerleider and Haxby (1994).

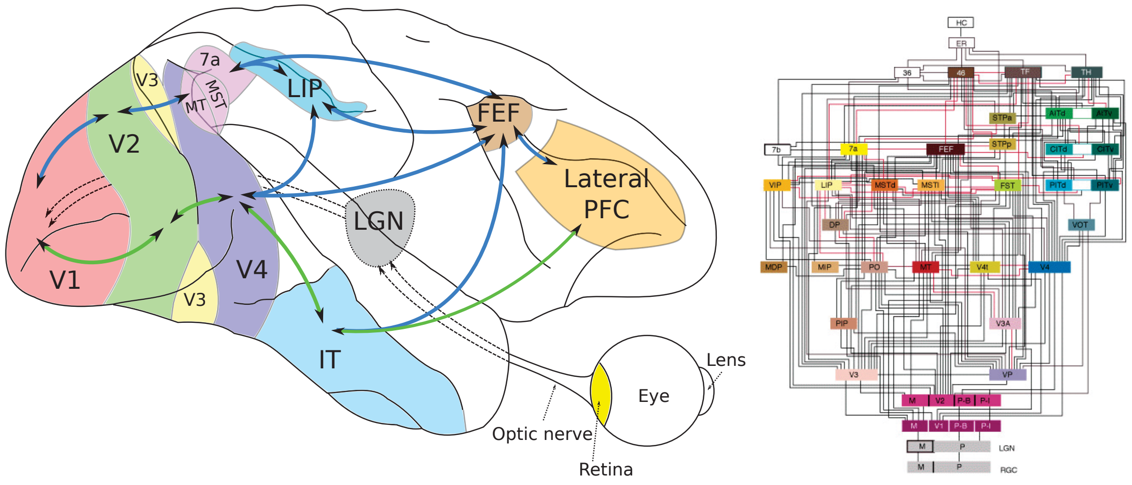
Visual attention can be seen as a top-down control of the visual system (Kastner and Ungerleider, 2000; Miller and Buschman, 2013). This control emerges from high-level cortices in the frontal lobe, especially from the FEF and the lateral PFC. It modulates the visual processing with spatial information from the FEF, or with task-relevant information from the lateral PFC.

These top-down connection and information flows are typically associated with top-down visual attention, on which this thesis focuses. Bottom-up visual attention may be processed over the shown bottom-up pathways for sensory information, or may utilize other pathways over subcortical structures which are also existing and have to be reported as involved in visual attention (for example via the superior colliculous, Krauzlis et al. (2013)).

After illustrating the system as a whole, we characterize the function of the involved cortical areas (Fig. 2.2a):

The primary visual cortex (V1) encodes very local and simple features like edges (Hubel and Wiesel, 1962, 1968), disparities (Read and Cumming, 2006), motion (DeAngelis et al., 1995), or color contrasts (Gegenfurtner, 2003).

The secondary visual cortex (V2) encodes combinations of V1 features like corners (Hegdé and Van Essen, 2003, 2007).



(a) (b)

Figure 2.2.: a) Schematic sketch of the visual attention system in humans. The sketch is adapted from Jamalian and Hamker (2016) and was refined according to the anatomical data of Felleman and Van Essen (1991), Grefkes and Fink (2005), and Ungerleider et al. (2008). Additionally, the areas 7a, LIP, and PFC were added as they are involved in visual attention too (Tsotsos et al., 2005; Miller and Buschman, 2013). Arrows denote the ventral (green) and dorsal (blue) streams of the visual cortex plus their connections to frontal cortices (FEF and lateral PFC). The areas are explained in the main text. The streams' primary functions are to process the type (green, ventral) and location (blue, dorsal) of an object (Ungerleider and Haxby, 1994). Bottom-up processing is denoted by arrows from left to right, top-down processing by arrows from right to left. The top-down connections mediate attention signals (Miller and Buschman, 2013). **b)** The actual wiring of the visual system is much more complicated. Adapted from Felleman and Van Essen (1991), showing the wiring in a macaque monkey.

The fourth visual cortex (V4) encodes complex shapes (Cadieu et al., 2007; Hegdé and Van Essen, 2007; Pasupathy and Connor, 2002).

The inferior temporal cortex (IT) encodes objects (Kriegeskorte, 2009; Op de Beeck et al., 2001; Serre et al., 2007a; Tanaka, 1996).

The third visual cortex (V3, Tootell et al. (1997)), *the medial temporal cortex (MT*, Born and Bradley (2005); Treue and Maunsell (1996)), *the medial superior temporal cortex (MST*, Duffy and Wurtz (1997)), and *the Brodmann area 7a* (Siegel and Read, 1997) are involved in motion processing (Tsotsos et al., 2005).

The frontal eye field (FEF, Heinzle (2006); Pouget et al. (2009)) and *the lateral intraparietal cortex (LIP*, Bisley and Goldberg (2010); Ninomiya et al. (2012); Steenrod et al. (2013)) are areas involved in spatial information processing and spatial attention. These aspects include functions such as encoding of spatial information, spatial attentional processing, transforming of information between different coordinate systems (along with attention pointers), and planing of eye movements.

The prefrontal cortex (PFC, Miller and Cohen (2001)) is responsible for high-level cognitive functions. Mainly relevant for attentional processing is its lateral part, which contains cells encoding task (Sakai, 2008) and category related information (Ashby and Spiering, 2004; Freedman et al., 2001; Seger, 2008; Seger and Miller, 2010). Yet, this lateral PFC is involved in many other functions than these relevant ones, for example, the dorsolateral part of the PFC facilitates working memory functions (Ashby and O'Brien, 2005; Vitay and Hamker, 2010; O'Reilly and Frank, 2006).

This brief overview depicts only areas which are very strikingly involved in visual attention. There exist many other areas in the brain also linked to visual attention, which are left out for clarity in Fig. 2.2a. First of all, top-down visual attention is only a part of the attentional processing, next to the bottom-up processing. The latter may run over the in Fig. 2.2a. shown connections and thus outlined areas, or as mentioned before over other networks for example evolving around the superior colliculus (SC, Heinzle (2006), Krauzlis et al. (2013)). The SC is a subcortical structure in the brain next to the thalamus and is involved in bottom-up attention (Krauzlis et al., 2013). Besides this general remark about bottom-up attention, we like to give some examples of visual attention in other areas. To start with, we like to discuss the role of the superior colliculus deeper. The SC is often thought to be involved in visual attention, and not only in bottom-up, but also in top-down. The SC has, among others, the role of executing eye movements. When finally the attentional processing signals direct the eye movement planning in spatial areas like the FEF, these structures and the visual cortex interact with subcortical structures like the SC to modulate the initiation and coordination of eye movements, and then the SC drives the eye movement by sending a command to the motor area of the eyes. However, according to the review of Krauzlis et al. (2013), which especially reviews the SC's role regarding attention, the SC seems to be predominantly involved in these motor consequences of visual attention. Furthermore, visual attention is also to be reported to affect for example thalamic nuclei (Miller and Buschman, 2013). Thalamic nuclei are parts of the thalamus, a brain structure located deep under the cerebral cortices and centrally in the brain (in Fig. 2.2a, the nuclei would be next to the "LGN"). One important nucleus in this context is the pulvinar. Cortical areas are typically connected towards and backwards to thalamic nuclei, hence attentional processing could theoretically operate also over thalamic connections instead over the direct connections between cortical areas. However, it is currently not clear to what extent this happens.

Effects of visual attention can also be observed in areas that are normally attributed to other forms of attention, because other forms of attention can interact with visual attention and thus their areas are sometimes reported as affected by visual attention too. Examples may be areas related to attention in other sensory domains like auditory attention (Fritz et al., 2007), areas involved in emotional attention like the amygdala (Liebold et al., 2015; Pessoa and Adolphs, 2010), and areas related to attention in the execution of behavior like the basal ganglia (executive attention, Posner and Rothbart (2007)). Finally, effects of visual attention can also be found in areas involved in motor commands outside the visual domain, like for grasping or body movements (Chinellato, 2008). Such areas are located in the parietal cortex, and can be affected under certain conditions by visual attention, for example when grasping an attended object.

To conclude, we give in this section a brief overview of the brain structures involved in visual attention. With the knowledge about these structures, the reader can better understand the neuronal processing of visual attention.

2.3. Neurophysiological findings and related models

Visual attention modulates the response of cortical neurons in a variety of different ways. Reynolds and Heeger (2009) provide a high-quality overview of possible effects, which we will sum up and extend in the following.

One of the first neurophysiological effects of visual attention was observed by Duncan and colleagues in the early 90s, and coined by them as "biased competition" (Desimone and Duncan (1995), also Desimone (1998); Reynolds et al. (1999); Lee and Maunsell (2010a)). We display the phenomenon's experimental setup in Fig. 2.3a as it is typical for neurophysiological attention experiments, and thus we would like to explain the important terminology on it. In this setup, two stimuli are placed in the receptive field of a recorded neuron. The *neuron's receptive field (RF)* designates the spatial area in which the neuron is driven by stimuli. From the two stimuli in the RF, one is chosen to drive the recorded neuron *(preferred stimulus)*, and the other to not elicit a response *(anti-preferred stimulus)*.

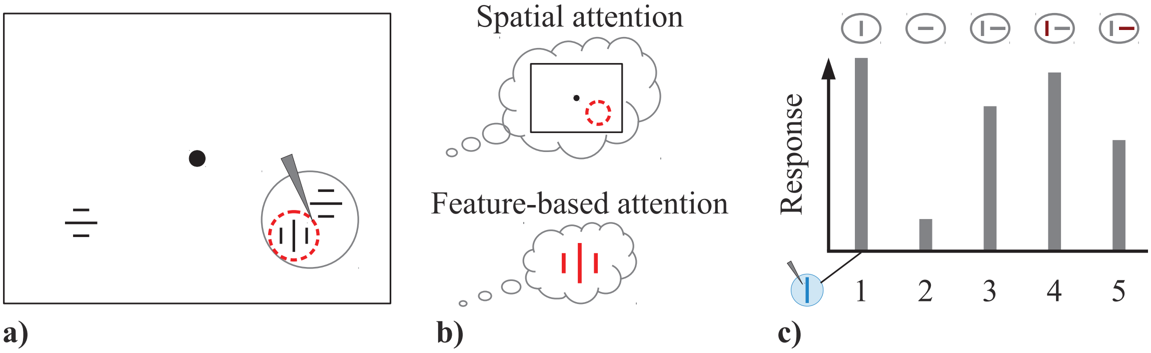


Figure 2.3.: a) The biased competition paradigm as a typical example to investigate the neurophysiological effects of attention. In this paradigm, a monkey has to fixate in the middle of the screen (black dot), while a neuron in the right hemifield is recorded. The neuron's receptive field (RF), i.e. the spatial area in which the neuron is driven by a stimulus, is illustrated as a gray circle. Two different stimuli are placed in the RF, one for which the neuron reacts (preferred stimulus, here a vertical grating), and one for which the neuron does not react (anti-preferred stimulus, here a horizontal grating). Attention is either directed to one of the two stimuli (red circle), or directed away from the RF. The latter is achieved by directed attention to a third stimulus outside the RF, here placed in the left hemisphere. **b)** Attention in biased competition can be spatial or feature-based. Spatial attention implies that the monkey is instructed to attend to a specific location like the one marked with the red circle, while feature-based attention implies to attend to a specific feature/object like the vertical grating. **c)** Attentional modulation of the neuronal response in biased competition (data adapted from Lee and Maunsell (2010a)). The figure shows the response of the neuron preferring the vertical grating in the following conditions: 1+2) presenting either the vertical or horizontal grating alone, 3) presenting both stimuli with attention directed away, and 4+5) attending to either the preferred or anti-preferred stimulus. The here shown firing rates are averaged values over the late response to a stimulus, as this period is typically modulated by attention while the initial peak is not (50 - 200ms after stimulus onset, as in the original paper).

Attention can be spatial or feature-based in biased competition (Fig. 2.3b), and can be either directed to one of the two stimuli *(attend preferred stimulus* or *attend anti-preferred* respectively), or away from them *(attend away)*. When attention is directed away, the finding is that the recorded neuron expresses a lower response when both stimuli in the RF are presented together than when the preferred stimulus would be presented alone (Fig. 2.3c). Thus, the two stimuli suppress each other, or they compete as coined by Desimone and Duncan (1995). Attention biases this competition in favor of the attended one: When attending to the preferred stimulus, the neuronal response increases, whereas when attending to the anti-preferred stimulus, it decreases. Thus, attention amplifies the response of neurons preferring the attended stimulus and suppresses the response of neurons preferring the un-attended stimulus.

However, this competition occurs only with different stimuli, similar ones do not elicit a suppressed response (Lee and Maunsell, 2010a; MacEvoy et al., 2009). Besides competitive effects, attention also alters the suppression from a stimulus beyond the receptive field (surround suppression, Sundberg et al. (2009)). Moreover, attention modulates *the contrast response function of a neuron (CRF*, Albrecht and Hamilton (1982)). The CRF describes the neuronal response when the excitation of the neuron is systematically increased, for example by increasing the contrast of its preferred stimulus. In some studies, attention amplifies the response prominently at lower contrast (contrast gain, Reynolds et al. (2000)), whereas others reported amplification at all contrasts (response gain, Williford and Maunsell (2006)). Attention further sharpens the neuronal tuning curve in the case of feature-based attention (Martinez-Trujillo and Treue, 2004), or scales it in the case of spatial attention (McAdams and Maunsell, 1999). These phenomena were reported by many other visual studies (Chap. 3), we have listed here only the most popular ones. In chapter 3, we will present a novel microcircuit model of attention replicating all these effects. Aside from these findings, attention can also modulate the spatial extent or shape of receptive fields (Anton-Erxleben et al., 2009; Anton-Erxleben and Carrasco, 2013), can modify the pattern of neuronal oscillations (Jensen et al., 2007; Grossberg and Versace, 2008), can decrease transmission delays (Sundberg et al., 2012), or can decrease the variability of the neuronal response (Mitchell et al., 2007). Attention has also temporal effects: attentional modulation does not occur during the response peak evoked by stimulus onset (Chelazzi et al., 1998), and attention can enhance the temporal coupling of neurons (Gregoriou et al., 2009). These studies show that attention modulates the neuronal activity in a rich set of ways.

Quite a lot of the physiological data is explained by existing single-area models (Boynton, 2009; Buia and Tiesinga, 2008; Hugues and José, 2010; Lee and Maunsell, 2009, 2010a,b; Ni et al., 2012; Reynolds and Heeger, 2009; Spratling, 2008; Wagatsuma et al., 2013). We will give in Chap. 3.4.1 a detailed discussion about the explanatory capabilities of each model. Many of these models use the idea of normalizing responses with a divisive signal (divisive normalization, Carandini and Heeger (2012)). Some proposed simple non-neuronal approaches constituting this idea (non-neuronal models, Fig. 2.4a): Boynton (2009), Lee and Maunsell (2009), Lee and Maunsell (2010a,b), Ni et al. (2012), while others presented neuronal models based on temporal-mean rate coded neurons (Gerstner

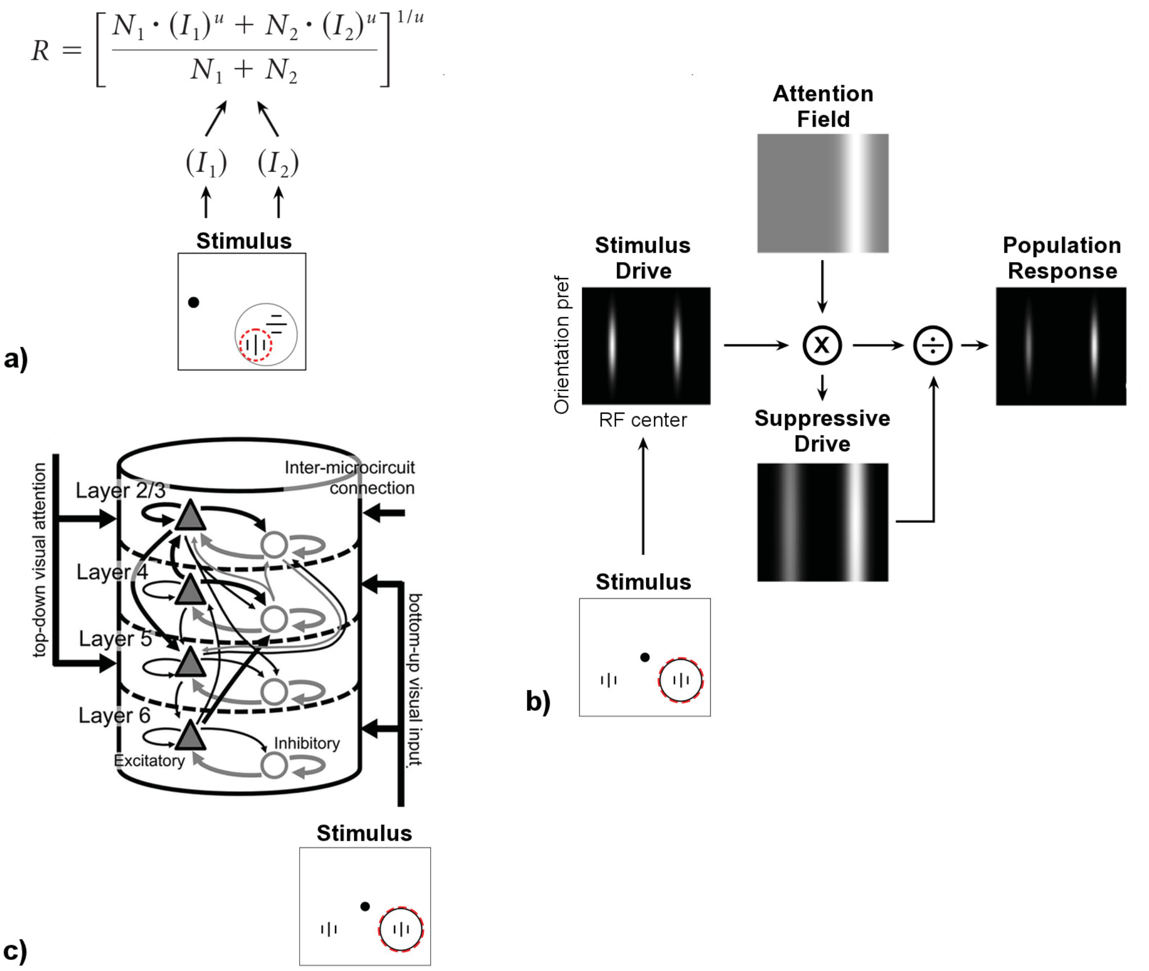


Figure 2.4.: Models explaining the neurophysiological effects of attention. **a)** The simplest model branch covers non-neuronal models, for example Lee and Maunsell (2010a). Their model has been designed to explain the interactions of stimuli within a receptive field of a neuron (RF). In the here-shown setup, the RF contains two stimuli, and attention is directed to the stimulus with the red circle. In the model, the two stimuli are represented simply by two variables (I_1, I_2). The model explains the response R, resulting from the interaction of both stimuli, by the shown equation. The equation integrates the inputs nonlinear, and normalizes the response by two divisive signals N_1, N_2. Attention affects these signals N_1, N_2, and in such way modulate the response. **b)** The next branch are models with mean-rate neurons, here shown exemplary on the popular model of Reynolds and Heeger (2009). A simple setup is illustrated with two stimuli, from which the right one is attended. In the model, the stimuli are represented by a neuronal input layer, the stimulus drive. This layer is then multiplied by a neuronal layer representing the attention signal (attention field), and divisively normalized by a suppressive drive. The neurons in each layer are organized according to their receptive field center (horizontal axis) and preferred orientation (vertical axis). **c)** The last branch comprises models with spike-rate neurons. We depict here the model of Wagatsuma et al. (2013) as it has a strong neuroanatomical foundation due to its simulation of the cortical microcircuit. All images were adapted from the original publications.

and Kistler, 2002, Chap. 1.5). A representative of the latter category is the powerful and popular model of Reynolds and Heeger (2009) (Fig. 2.4b). As our microcircuit model (Chap. 3) will be based on Reynolds and Heeger (2009), we will explain their model and the invoked concepts more deeply. Their model contains four layers: a stimulus drive E, an attention field A, and a suppressive drive S, and the neuronal response R. Each layer is a 2D population of neurons, organized according to their receptive field center and preferred feature (e.g. orientation). The core equations of the model are depicted in Eq. 2.1, 2.2.

$$R = (E \cdot A)/(\sigma + S) \tag{2.1}$$

$$S = s * (E \cdot A) \tag{2.2}$$

Whereby σ denotes the contrast gain of the neuron's response, s is a 2D Gaussian kernel representing local inhibitory connections, and the symbol * stands for convolution.

In the model, the stimulus is represented by the stimulus drive (E), while the attention field (A) represents the attended location or stimulus feature. Neurophysiologically, attention enhances multiplicatively the neuronal responses. In the model, this effect is simulated by multiplying E with A, whereby $A = 1$ without attention and $A > 1$ with attention (Eq. 2.1). The outcome is divided by the suppressive drive to obtain the neuronal response. This concept is known as *divisive normalization*, which has the aim to normalize the neuronal response based on suppressive influences (Carandini and Heeger, 2012). Divisive normalization is widely observed in primates, and explains in attention setups the modulation of the contrast response function. The normalization depends in this model on the influences incorporated into the suppressive drive (Eq. 2.2): the stimulus drive, the attention field, and local inhibition between features or spatial locations (s). Inhibition effects are often observed in neurophysiological attention data sets, for example in biased competition. By these mechanisms, the model can explain a moderate range of neurophysiological attention effects: feature-based biased competition, attentional modulation of the contrast response function (both response and contrast gain), and attentional modulation of the neuronal tuning curves (both scaling and sharpening). In our model, we will use similarly the concepts of amplification, normalization, and suppression, but we will develop them further in terms of a neuroanatomical foundation.

However, none of the aforementioned models can explain all data sets (a detailed discussion will be given later in Chap. 3.4.1), and they lack a neuroanatomical proposal how attention can be implemented within the cortical microcircuit (Douglas and Martin, 2004). A few other models have replicated at least some of the connectivity of cortical microcircuit, but cannot explain many of the data sets. Spratling (2008) developed another model based on temporal-mean rate coded neurons, while Buia and Tiesinga (2008), Hugues and José (2010), Wagatsuma et al. (2013) propose models with spiking neurons (Fig. 2.4c). The dichotomy between sophisticated model abilities and a strong neuroanatomical foundation is not surprising as modeling the cortical microcircuit greatly increases the complexity of the model, which either prevents to model a particular experiment or increases greatly the effort to simulate many experiments. Aside from single-area models, a few of the previously-outlined physiological data sets were also replicated by neuro-computational system-level models of attention (Bayerl and Neumann, 2007a; Chikkerur et al., 2010; Corchs and Deco, 2002; Hamker, 2004b, 2005a; Hamker and Zirnsak, 2006; Hamker et al., 2008; Lanyon and Denham, 2004, 2009; Rothenstein and Tsotsos, 2014). These models will be discussed in section 2.5.

In summary, many models already exist to explain the neurophysiological effects of attention, however many questions remain open. First of all, the neuronal mechanisms of attention need clarification as no model can explain all the data sets with its mechanisms. This question will be addressed by our novel microcircuit model in Chap. 3. Furthermore, models with temporal-mean rate neurons lack a neuroanatomical proposal of how attention might be implemented in the cortical microcircuit (Douglas and Martin, 2004). Neither it is clear which connections in the circuit transports attention signals, nor how basic neuronal effects like amplification, normalization, or suppression are implemented. Some attempts already exist, for example Brosch and Neumann (2014) propose an implementation of the amplification via interneurons. Hence, the mean-rate neuron models should be further developed in terms of their neuroanatomical substrate. We will contribute to this aim by proposing a concrete implementation of the suppression mechanisms in the microcircuit model.

2.4. Visual attention in object localization and recognition systems

This section illustrates the different approaches of how visual attention is used in object localization systems. Object localization is defined as the task to search a given target in a scene, thus a system has to search, to localize, and to recognize the target object. Hence, many localization systems can also do a recognition, and due to this reason, they are sometimes called recognition systems in the literature. The task is daily executed by humans, for example when meeting a person in a crowded scenery or picking up a pen at an office desk. We will distinguish models according to their processing in three approaches: 1) saliency models using bottom-up attention as a pre-selection stage, 2) saliency models with top-down attention, and 3) models that are holistically controlled by top-down attention.

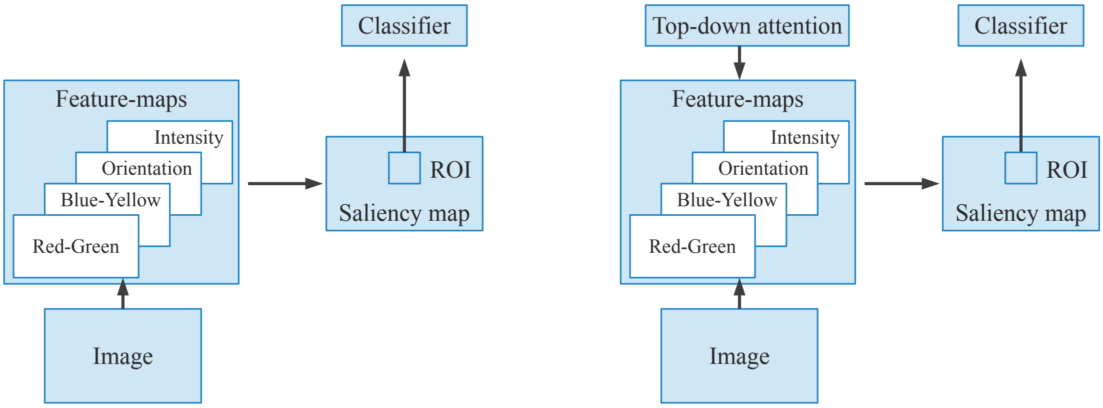
Psychologists investigate the object localization task in the paradigm of guided visual search (Wolfe, 1994), in which a given target has to be searched among distractors. The search can be conducted via eye movements. The eye movement itself is called a saccade, whereas focusing with an eye is called a fixation. The task performance is measured via reaction times (RTs), or via the chain of fixation points (scan-path). The scan-path can also be simulated by saliency models, allowing to compare them with human performance (Borji and Itti, 2013). If the target differs from the distractors via a simple feature, it is easy recognizable (pop-out effect, Wolfe (1994)), resulting in a low RT. As the RT in pop-out scenarios is typical independent from the number of distractors, it indicates that subjects search trough the scene in parallel (parallel search). If the target is more similar to the distractors, the search is slower. A popular instance of this case is conjunction search, in which the target is defined by the conjunction of two simple features, whereby one of the two features is part of the distractors too. A typical example is searching a green-vertical bar among green-horizontal and red-vertical bars. Subjects typically fixate several items until they found the target in conjunction search, thus the RT increases proportional to the number of distractors, denoted as serial search. Attention is deployed to each fixated item in the task, thus it indicates the selection of potential target items (overt attention). However, attention can also be deployed to locations without eye movements (covert attention, e.g. Posner (1980)), thus the task can also be performed without eye movements, leading similarly to the deployment of attention to potential target locations (Wolfe, 2000).

Multiple comprehensive reviews exist about object localization systems. They provide different foci: Frintrop et al. (2010) present a good general overview of the topic, but does not cover the most recent developments and holistic top-down attention models. Hamker (1999) and Frintrop et al. (2010, Sec. 2.3) review the psychological theories that have inspired the saliency models: the feature integration theory of Treisman and Gelade (1980), the early saliency model of Koch and Ullman (1985), and the guided visual search 2.0 model of Wolfe (1994). Borji and Itti (2013) give a high-quality and condensed survey which compares 52 different saliency models. Filipe and Alexandre (2013) is very similar to Borji and Itti (2013), but it is more general written. Jamalian and Hamker (2016) concentrate on robotic setups and the role of spatial processing.

2.4.1. Saliency models with pure bottom-up attention

Saliency models localize objects by using attention as a spatial pre-selection mechanism for a subsequent, sophisticated object classifier (Borji and Itti, 2013; Frintrop et al., 2010). In the bottom-up approach (Fig. 2.5a), the idea is that attention emerges in a bottom-up manner from the scene and represents locations with large amount of information, called conspicuity. These conspicuities are integrated to a saliency map representing salient locations, from which one location is selected for the classifier. The conspicuities are typically extracted by filtering the image with local feature detectors, resulting in several feature maps. Thus, these maps are then integrated to the saliency map. Locations are selected in form of Region of interests (ROIs) around the highest saliency values, and finally the best ROI is passed to the classifier to recognize the object in this region. If the target object is not in this region, the ROI with the next lower saliency value is explored by inhibiting the previous location (inhibition of return, IOR). This process iterates until the target is found. The main advantage of this approach is the high execution speed as only a few image regions have to be processed by the computational expensive classifier (Frintrop et al., 2010). Disadvantageously, the approach can only localize salient stimuli, and it is unknown how to integrate optimally the feature-maps into a saliency map (Frintrop et al., 2010).

A typical example is the Neuromorphic Vision Toolkit (NVT) from the group of Itti (Itti et al., 1998; Itti and Koch, 2001). They use local feature descriptors, which encode red-green color contrasts, blue-yellow color contrasts, illumination, and local orientation. The saliency map is obtained by a normalized linear-combination of these feature maps (Itti et al., 1998). The framework is very popular as its source code is freely available, and as it is maintained and further developed by the group of Itti. Besides the NVT, there are many other studies describing similar systems or applications to specific domains. For example, many systems have been applied to real-world problems: Aryananda (2002), Frintrop et al. (2004), Miau et al. (2001), Ouerhani (2003), Salah et al. (2002), Sun and Fisher (2003), Vincent et al. (2007), Walther et al. (2004, 2005), and the popular proto-object approach of Walther and Koch (2006). An overview of the existing models can be found in the high-quality review of Borji and Itti (2013). It lists 52 saliency models of which 39 belong to the bottom-up approach. Borji and Itti (2013) present them in a very compressed form, while more details can be found in the review of Filipe and Alexandre (2013).



(a) Bottom-up saliency model

(b) Top-down saliency model

Figure 2.5.: a) Approach of bottom-up saliency models, e.g. the Neuromorphic Vision Toolkit (Itti et al., 1998). Such a model uses attention as a spatial pre-selection stage for a subsequent, sophisticated classifier. The image is processed by several maps encoding simple features (feature-maps), which then are integrated towards a map representing salient locations. From them, one is selected (region of interest, ROI) and passed to the classifier for recognizing the object in this ROI. **b)** Approach of top-down saliency models, e.g. VOCUS (Mitri et al., 2005). Such a model operates similarly as in (a), except that top-down attention is deployed to the integration process of the feature-maps to the saliency map. The top-down attention encodes features belonging to the target object, thus the resulting saliency map contains potential target locations and not only salient locations. The figures in (a) and (b) are drawn based on the model description given in the original publications of Itti et al. (1998) and Mitri et al. (2005) respectively.

2.4.2. Top-down guided saliency models

More sophisticated systems use top-down attention to bias the saliency map in favor of target objects (Fig. 2.5b), so that the region of interests (ROIs) contain primary potential target objects. The idea dates back to the guided visual search model of Wolfe (1994), i.e. that attention guides vision for particular objects. Relevant models of this category are reviewed by Frintrop et al. (2010), Borji and Itti (2013), Filipe and Alexandre (2013). In the following, we will present the approach at the example of the popular model VOCUS, and we will give an overview of advantages, successfully applications in real-world scenes, and variations of the idea.

The model VOCUS from Frintrop (Mitri et al. (2005), also Frintrop et al. (2004), Frintrop (2006)) can deploy top-down attention to the low-level feature maps. The attention signal represents a modulation weight, which weights each feature-map during the integration to the saliency map. The weight encodes if a particular low-level feature is part of the target object. Thus, the resulting saliency map is biased for the target objects, and the regions of interest (ROIs) contain mostly potential target objects. Afterwards, the best ROI is passed to a classifier to recognize the target object. For example, Mitri et al. (2005) use as classifier a classification and regression trees (CART) with Ada-boost (Viola and Jones, 2004; Freund and Schapire, 1996). The target objects are typically task-dependent, thus the top-down attention signals are created task-dependent too. In VOCUS, the top-down weights are learned in a preceding learning phase (Frintrop, 2006). In it, the user marks at first the region of the target as ROI, and then, the features maps and the most salient region (MSR) within these ROIs are calculated, as well as the feature maps in the remaining image. From these information, the weights are learned so they distinguish at best the target from the rest of the image. Mitri et al. (2005) use this system to recognize the soccer ball in a RoboCup soccer scenario. The learned weights for the ball deploy attention prominently to the color feature 'red', as the ball was marked red and could so be well distinguished against the green playing field. This concept of a top-down guided saliency map is used similarly in other models: Borji et al. (2009, 2010), Navalpakkam and Itti (2006), Paletta et al. (2005), Xu and Chenkov (2009).

The top-down saliency approach has the advantages that it recognizes non-salient objects and that it greatly eliminates false positives. Non-salient objects are not recognized in the bottom-up approach as the saliency map does not represent them (Frintrop et al., 2010). False positives are greatly reduced as shown by Mitri et al. (2005). The classifier without attentional pre-selection found the target in 146/180 test scenes, but reported also 160 false positives. When attention was used, the system classifies objects only at a few potential target locations, which greatly reduce the false positive rate from 160 to 23 and does not impair the recognition rate (from 146 to 141). Besides these benefits, the top-down saliency approach speeds up the processing like in the bottom-up saliency models. Due to these advantages, the approach has been very often and successfully used in real-world and robotic applications: Borji et al. (2009), Frintrop et al. (2004); Frintrop (2006), Gao and Vasconcelos (2005a,b), Li et al. (2010), Mitri et al. (2005), Navalpakkam and Itti (2005), Oliva and Torralba (2003) , Rasolzadeh et al. (2009), Torralba (2003b).

The concept can also be realized via a Bayesian modeling approach (Congdon, 2007; Edwards et al., 1963). In Oliva and Torralba (2003), the system calculates the saliency as the probability that the target object is present at a certain location. The saliency probability depends on the low-level features, and on the top-down attention implementing a probabilistic prior. The top-down prior is learned in a preceding learning phase via Bayesian inference, i.e. they calculate for each low-level feature the probability to predict a certain object class, and afterwards calculate the prior from this probability via the Bayes theorem. The idea of modeling top-down influence via a Bayesian framework is used by many models: Elazary and Itti (2010), Gao and Vasconcelos (2005a,b), Li et al. (2009, 2010), Oliva and Torralba (2003), Pang et al. (2008), Sprague and Ballard (2004), Torralba (2003a,b).

A sophisticated variation of the saliency models are proto-object models (Walther and Koch, 2006; Jamalian and Hamker, 2016). They attend to proto-objects that are a region of grouped features, potentially belonging to an object. The idea is inspired from the finding that attention can be deployed to whole objects (Roelfsema et al., 1998). Attention serves again as a spatial pre-selection mechanism, hence it selects a region shaped as a potential object. This region is better adapted to the size and the shape of objects as the standard approach of a rectangle-shaped region of interests. It has the advantage that only relevant information is passed to the subsequent classifier, potentially increasing its performance. Furthermore, the proto-object gives access to the object features at the selected region. A popular example is the model of Walther and Koch (2006). The model obtains at first a saliency map like in the classical bottom-up saliency models. Top-down connections transport then the most salient location back to the feature-maps, and the feature-map is determined which contributes at most to the saliency. In this winning feature-map, attention is spread around the most salient location, grouping features together. This results in an object specific region, representing the proto-object. The following models obtain also proto-objects (Jamalian and Hamker, 2016), although their grouping strategies differ a lot: Palomino et al. (2011), Russell et al. (2014), Walther and Koch (2006), Yanulevskaya et al. (2013), Yu et al. (2008), Yu et al. (2010), Yu et al. (2013).

The concept of top-down saliency models is varied further by several studies: Björkman and Eklundh (2006), Fritz et al. (2005), Navalpakkam and Itti (2005), Peters and Itti (2007), Rasolzadeh et al. (2009), Rybak et al. (1998), Wischnewski et al. (2010). One interesting variation is the idea of a task-relevant map, which highlights all task-relevant locations (Navalpakkam and Itti, 2005). In contrast to the biased saliency map, this map encodes high-level cognitive information. Like in the classical version, top-down attention is projected to an early processing stage with low-level features filters and a biased saliency map is obtained from this. However, the model obtains also other information like scene gist or scene layout from the low-level features. This information is used together with working memory and long-term memory contents to obtain a task-relevant map (TRM).

This map contains locations of objects that might be relevant for the current task. The locations are determined by the scene gist and scene layout in combination with the stored objects in memory. By this, the map represents a high-level cognitive expectation of relevant object locations for the current task.

Another powerful variation is the active, robotic vision system from the group of Björkman (Rasolzadeh et al. (2009), also Björkman and Eklundh (2006)). They combine several engineering concepts together to create a model capable of detecting, fixating, and manipulating objects in the real world. Concerning the vision part, they have separate modules for attention, segmentation, and object recognition. They improve the classical top-down saliency approach by using two saliency maps, a bottom-up and a top-down, and combine them dynamically via an information entropy measurement. From the resulting map, they select regions of interest (ROIs) and fixate the best ROI by the binocular camera system. The fixation reduces disparity from the binocular vision, and increases the resolution of the inspected object due to a second camera pair realizing a fovea, both simplifying the segmentation and the object recognition tasks. After fixating, segmentation is performed based on motion and disparity cues. Finally, objects are recognized via color histograms and SIFT. SIFT encodes an object via an unordered set of key points based on local maximum of Gaussian filters (Lowe, 2004). The overall performance of the system is quite high, and also the combined saliency map performs quite-well as the first ROI contains the target object in 78% of the tests. The result was measured with 10 different objects and 96 real-world test scenes.

In summary, top-down saliency models represent a powerful and broadly used approach. On the other hand, the approach is not without disadvantages, for example it lacks biological plausibility. We will discuss these disadvantages in the next section as the holistically attention models will solve them.

2.4.3. Models with top-down attention as holistic, cognitive control

In the most sophisticated category, top-down attention modulates neuronal activity in the whole system, thus we denote its influence as holistic. Attention signals project top-down from areas encoding object-categories to high-level visual areas, and afterwards to low-level visual or spatial areas. Hence, attention controls both the recognition processes in

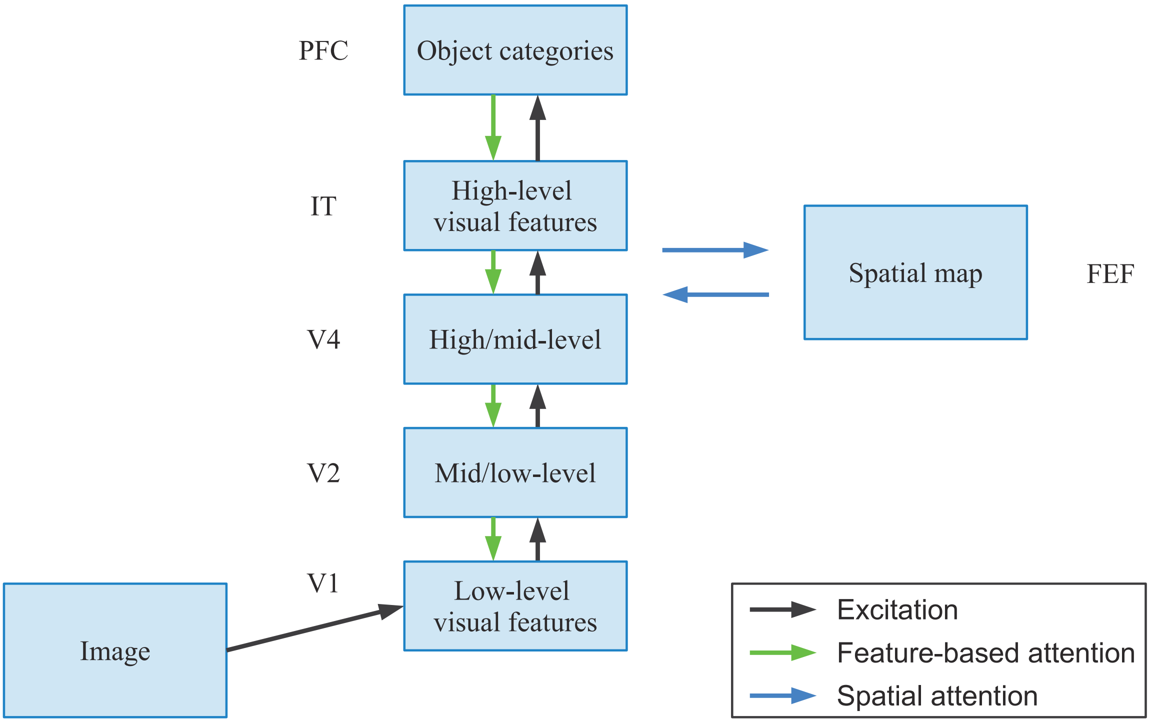


Figure 2.6.: Models with top-down attention as holistic control, e.g. Hamker (2005b). In such a model, the image is processes by a hierarchy of visual areas that encode increasingly more complex features, ranging from edges in the lowest area to object views in the highest area. At top of the visual areas, an area encoding object categories is situated. Attention serves as a process to control the model for the current task, i.e. object localization. This task is performed by deploying attention to the target object, conducted by activating a target-object encoding cell in the highest area. From this area, top-down attention signals are deployed firstly to the highest visual area, and afterwards to lower visual and spatial areas. These signals amplify target relevant features in each of the visual areas (feature-based attention), and locations in the spatial areas (spatial attention). Other features and locations are suppressed. At the end of this processing, the spatial map denotes the location of the target. Therefore, attention modulates the neuronal activity in whole model, leading us to term its influence as holistic. The figure is based on the attentional processing in the human brain. A particular model may contain fewer areas as some cortical areas may be left out for simplicity (see model descriptions in the main text). In the figure, the cortical areas are abbreviated as follows: V1, V2, V4: primary, secondary, and fourth visual cortex; IT: inferior temporal cortex; PFC: prefrontal cortex; FEF: frontal eye field.

the high-level visual areas and the selection or localization processes in the spatial areas. Therefore, we describe attention in this category as a cognitive, holistic control process. Fig. 2.6 shows the general structure of models with such attentional processing.

This attentional processing is what happens in the primate brain, thus many neuro-computational system-level models of attention could be put into this category (Sec. 2.5). However, the majority of them have been designed for setups with simple stimuli like red or green bars, whereas we are interested in real-world object localization scenarios. Thus, we put only models in this category that are capable to handle whole objects, and have been shown to work with real-world scenes. Besides neuro-computational models, this category contains also two models from the computer vision community (Walther and Koch, 2007; Chikkerur et al., 2010), which are not closely linked to brain areas or psychophysical attention data. In the following, we will explain the existing models of this approach before we elucidate its advantages. The proposed model will also fall into this category, hence we will cover the approach and its models in deep.

The first three models, Antonelli et al. (2014), Beuth et al. (2010), Hamker (2005b), belong to Prof. Hamkers research group and are neuro-computational system-level models capable of object localization and recognition. We will explain at first the model of Hamker (2005b) and afterwards list the improvements in Antonelli et al. (2014), Beuth et al. (2010). The model of Hamker (2005b), which will serve as the basis for our novel system-level model (Chap. 4), contains two low-level visual stages comparable to an area like V1, two high-level visual stages 1 and 2 representing the areas V4 and IT, two spatial stages representing the frontal eye field (FEF), and an object stage representing the prefrontal cortex (PFC). The model is described also in Hamker (2005a) from a neuroscientific point of view, thus biological background, including the mapping from stages to cortical areas, can be found there. The low-level stages simulate an enhanced version of a saliency model (Itti et al., 1998), hence their neurons encode edges, color contrasts and illuminations. Areas V4 and IT pool these to a spatially more invariant representation of low-level features and modulate them with attention. The object localization task is realized by feature-based attention towards the target object, implemented by activating a particular object neuron in PFC. This activity results in a top-down attention signal to IT and V4 that amplifies all neurons encoding a feature of the target. In addition, inhibition within an area suppresses neurons encoding task irrelevant features or objects, e.g. distractors or background clutter. The attentional-modulated neuronal activity in the high-level visual areas is projected to the oculomotor system (FEF), biasing it to fixate the target location. The fixated object is classify via match units, and its location is inhibited in case of a non-target (IOR approach) which triggers the processing of the next potential target location. The FEF projects back to the visual areas via a reentrant signal, forming a recurrent loop. Spatial attention emerges from this loop, amplifying responses at the target location and suppressing others. The attention process mediates the holistic aspect of the control as modulates both the visual and spatial areas.

Beuth et al. (2010) and Antonelli et al. (2014) are preceding works of the author and present a single model in two applications: a virtual reality and a robot. The robot was developed in the frame of a joint European project with several partners, whereby the author contributed the object localization system. The system has also been published additionally in Beuth et al. (2014), to explain closer its amplification and suppression mechanisms. The author's model improves the work of Hamker (2005b) regarding its applicability to real-world scenes. For this purpose, the system has been simplified to make it more robust: The inhibition operates for example via top-down connections, the IOR approach has been removed as it was not necessary in these applications, and a comparison to neurophysiology has been omitted. But most importantly, the work improves the previous model with the ability to learn high-level object descriptors. Objects are represented by view-tuned cells, as found in high-level visual cortices (Logothetis et al., 1995), encoding a particular view of an object. An object view is formed by a pattern of low-level features, representing a spatial arrangement of edges and colors. Contrary, the model of Hamker (2005b) represents an object as a set of low-level features without spatial arrangement. The new view-tuned cells reside in a higher visual area (HVA), representing an abstract high-level visual area, comparable with the cortical areas V4 or IT. To learn the view-tuned cells, a method called trace learning is employed, which relies on the temporal continuity in the visual input stream (Földiák, 1991; Spratling, 2005). In this doctoral thesis, we will precisely integrate the new approach to learn object descriptors into our novel system-level model. It will allow us in Chap. 4 to learn the many different objects.

All three studies reported high localization accuracies, but used relatively easy tasks with a low number of objects and test scenes. Hamker (2005b) reported 81% accuracy to fixate the target object correctly in the first four fixations and 50% to fixate it immediately correctly. They used 16 different objects in a data set of three real-world scenes. Beuth et al. (2010) reported a 100% ability to discriminate correctly between 10 different objects in a virtual reality, but this evaluation was done on scenes containing only a single object. Antonelli et al. (2014) reported a 100% accuracy to localize 3 objects in 27 different indoor laboratory scenes via the first fixation. In summary, the reported accuracies are really high, but also the tasks were relatively easy due to low number of objects and test scenes.

Walther and Koch (2007) present an object recognition and localization system which is based on HMAX (Riesenhuber and Poggio, 1999; Serre et al., 2007a,b) and guided holistically by top-down attention. The model consists of the stages S1, C1, S2, C2, VTU, 'objects', 'categories', and 'task'. The first stage S1 extracts low-level visual features like edges from the image, stage C1 pools them, stage S2 extracts combinations of C1 features like combined edges (mid-level visual features), C2 pools them again, and view-tuned units (Logothetis et al., 1995) extracts views of an object from the C2 response pattern (high-level visual features). This structure is similar to the classical HMAX, but Walther and Koch (2007) added on top of it stages encoding: whole objects, the category of each object, and the current task. Top-down attention is deployed via feedback connections projecting downwards this hierarchy, hence via the chain 'task'→'categories'→'objects'→C2→S2. The feedback connections are reciprocal designed to the feedforward connections. The top-down signal biases the recognition process for the current task by selecting relevant categories, objects, VTU, and mid-level features in C2 and S2. The model is part of a larger "unifying framework for attention and object recognition (UNI, Walther and Koch (2007))". They use it to compare the different attention approaches of Serre's research group in one framework (Walther et al., 2005; Walther and Koch, 2006, 2007). For this purpose, the framework contains also a spatial processing. If an object should be searched in the scene, they extracted a saliency map from the S2 layer to calculate a fixation point. Instead of this, they can also attend to a particular location in the scene by deploying spatial attention to the system. However, they simulate neither spatial cortices nor the attentional processing in them. The system was evaluated at the task to detect frontal views of human faces. The best top-down attention configuration achieves an immediately correct fixation in 38% of the tests, but needs also more than three fixations in 49% of the tests. These results indicate a moderate performance of the system.

Another holistic model is the Bayesian inference theory of attention from Chikkerur et al. (2010). According to this theory, the primate brain has the goal to infer the identity and the location of a target object in parallel. Within this process, spatial attention reduces the uncertainty of the object identity by constraining it to the subset of objects at a particular location, and feature-based attention reduces in the same manner the uncertainty of the object location. Attention is implemented in the framework as prior information. This process could either increase or decrease the probability of a certain feature or location. The models contains the stages I (an abstract input stage like the image or V1/V2), X (correspondent to V4), F (to IT), L (to FEF or LIP), and O (to PFC). They have a feedforward processing via Input→V4→IT→PFC, a spatial attentional processing via FEF/LIP→V4, and a top-down processing via PFC→IT→V4. They can replicate three physiological attention effects and the pop-out effect from guided visual search (Wolfe, 1994). If the model was used at real world scenes, they preprocess with HMAX (Riesenhuber and Poggio, 1999; Serre et al., 2007b,a) the scenes and set the resulting C2 activations from HMAX as input I. The evaluation on a real-world data set, containing cars and pedestrians, shows that the system can localize cars with 82% and pedestrians with 81% accuracy. On the same set, a standard bottom-up saliency model (Itti et al., 1998) achieves 43% and 39% respectively. The high performance might be explained by their weak criteria for a successful localization, because they define an object as correctly localized if its center contains activity in a thresholded saliency map ($P(L|I)$).

The approach of attention as holistic, cognitive control has the following three advantages.

Firstly, it has the benefit that object recognition, localization and attention are closely intertwined, a goal already suggested by Frintrop et al. (2010). Attention operates no longer as a solely spatial pre-selection stage, instead it controls in parallel recognition and spatial selection. The parallel processing solves the interdependency problem of object recognition and selection, which was unsolved in the previous approach: Spatial selection needs knowledge about the properties of the target object and so depends on a successful recognition. However, recognition requires beforehand spatial selection to segment the object from the background (Antonelli et al., 2014; Beuth et al., 2010). Spatial selection is achieved in the models via localization in two slightly differently approaches. Chikkerur et al. (2010) and Walther and Koch (2007) amplify the features of the target in the visual areas, and the resulting activity blob indicates the target location. This strategy might fail if several target objects are present in the scene, as the system would amplify multiple locations. Antonelli et al. (2014), Beuth et al. (2010), Hamker (2005b) contain a model of the frontal eye field, focusing spatial activity on the target location. In case of multiple targets, attention would be initially split between the locations, but then would be focused on a single location (split of attention, Zirnsak et al. (2011a)). To sum up, the first advantage is that attention controls recognition and spatial selection in parallel, avoiding the problematic separation of both processes.

Secondly, the holistic approach allows to spatially select an object via a high-level visual representation instead of low-level features as in the top-down saliency models. These deploy attention directly to the low-level features, which allows a selection only via them. This concept is only suitable in some tasks, for example Mitri et al. (2005) utilized it to localize a red ball in a soccer scenario. They used the low-level feature 'red' as the ball color was uniquely in their scenario. The approach works also for localizing faces, as they can be well localized by the unique color hue of the human skin (Darrell et al., 2000). However in many tasks, the target objects cannot be selected uniquely by low-level features (Xu and Chenkov, 2009). Therefore, attention has to be deployed to a more complex visual representation. These range from mid-level representation (in area V4 or stage C2), view-tuned cells (in IT, HVA, or VTU), to object-category representations (in PFC, 'objects', or 'categories') in the holistic models (Antonelli et al., 2014; Beuth et al., 2010; Chikkerur et al., 2010; Hamker, 2005b; Walther and Koch, 2007). Spatial selection is then based on these representations. They are also utilized for the recognition and thus are shared between recognition and spatial selection. Walther and Koch (2007) denote this concept as feature-sharing.

Finally, the holistic approach has a much higher biological plausibility than the top-down saliency approach. Firstly, the attentional processing in the models resembles the cortical attention network (Sec. 2.2). The network projects an attention signal downwards from task-encoding cortices (PFC) over high-level visual cortices (IT) to mid- (V4) and low-level visual cortices (V2, V1). The chain is simulated in all models until mid-level cortices. Furthermore, the cortical attention network includes a second signal chain from the frontal eye field (FEF) to dorsal cortices (LIP, V2, V1), and to ventral cortices (V4, IT). All models simulate connections from a spatial stage to the ventral cortices, but only Antonelli et al. (2014), Beuth et al. (2010), Hamker (2005b) simulate them biological accurately in both directions. Moreover, only these models simulate a biological realistic frontal eye field. Secondly, biological plausibility arises from the consistency between attention to object-specific neurons and the psychological concept of object-based attention. As reviewed by Walther and Koch (2007), several studies (Duncan, 1984; Egly et al., 1994; Roelfsema et al., 1998) found that attention can be deployed to objects or groups of objects. They found that (spatial) attention is spread over an object which is defined by luminance contrast (Egly et al., 1994), color (Mitchell et al., 2003; Reynolds and Desimone, 2003), or illusory contours (Moore et al., 1998). This result was even observed when two objects were semi-transparent superimposed (O'Craven et al., 1999). To sum up, the high biological plausibility of the approach arises from resembling the cortical attention network and from consistency with object-based attention.

The only major disadvantage is that the approach has been recently developed, so there exist only a few models and these have not been thoroughly benchmarked with real world task and many objects classes. Models of our group have been used with 3 - 16 object classes (Antonelli et al., 2014; Beuth et al., 2010; Hamker, 2005b), whereas Chikkerur et al. (2010) and Walther and Koch (2007) uses two-class tasks. Furthermore, all these studies focus on the proof of the novel concepts, and hence might have not tuned their system for highest accuracies.

To conclude, the approach of attention as holistic, cognitive control might be the most sophisticated approach for object localization and recognition. The concept has several advantages over the top-down saliency models: 1) attention controls in parallel recognition and spatial selection, 2) it selects an object via sophisticated high-level visual representations, and 3) it has a high biological plausibility. On the other hand, the approach is implemented by only a few models and it lacks a rigorous testing with many object classes or in real-world scenarios. This motivates us precisely for our work in Chap. 4: a further development of the holistic models and a benchmark on a realistic localization task.

2.5. Neuro-computational system-level models

Visual attention is also modeled at the level of multiple brain areas, typically to explain either psychophysical phenomena or general principles of attentional processing in the brain. We regard a neuro-computational model as a system-level model if it contains multiple, connected brain areas, and is linked to neurophysiology and neuroanatomy. We also consider only recent models here, namely after the year 2000. A review of older models can be found in Hamker (1999) and in Tsotsos and Rothenstein (2011). A few neuro-computational models fit also in the holistic object recognition and localization category (Sec. 2.4.3) as they are capable to cope with real-world objects in guided visual search: Antonelli et al. (2014); Beuth et al. (2010); Hamker (2005b). Visual attention is involved in many psychophysical phenomena (Carrasco, 2011), a full list of them is illustrated in Tab. 2.1. In general, attention improves the behavioral performance in these tasks, yet the detailed effect depends on the phenomenon, thus we provided a list of references. Many of these phenomena can be simulated by system-level models of attention, either explaining the phenomenon fully as attentional phenomenon or investigating single attention effects in it. We conducted a literature search about the currently available models, and listed them, ordered per phenomena, in Tab. 2.1 too. From this list, it is visible that some phenomena have been never simulated: meta-contrast masking, object substitution masking, crowding, attentional blink, and multiple object tracking.

Psychophysical phenomenon	Reference for the phenomenon	Neuro-computational system-level models of visual attention
Visual search	Wolfe (1994); Lee and McPeek (2013)	Antonelli et al. (2014); Beuth et al. (2010, 2014); Bruce and Tsotsos (2011); Deco et al. (2002); Deco and Rolls (2004); Hamker (2004a, 2005a,b); Lanyon and Denham (2004); Rodríguez-Sánchez et al. (2007); Zaharescu (2004); Zirnsak et al. (2011a)
Meta-contrast and objects substitution masking	Breitmeyer and Ogmen (2006); Goodhew et al. (2013)	-
Crowding	Whitney and Levi (2011)	-
Change detection	Rensink (2002)	Hamker (2005c)
Spatial compression	Anton-Erxleben and Carrasco (2013); Zirnsak et al. (2014)	Hamker et al. (2008); Zirnsak et al. (2010)
Attentional blink	Shapiro et al. (1997)	-
Multiple object tracking	Cavanagh and Alvarez (2005)	-
Saccadic receptive field changes	Zirnsak and Moore (2014)	Hamker and Zirnsak (2006); Hamker et al. (2008); Fix et al. (2010); Ziesche and Hamker (2014); Zirnsak et al. (2010, 2011a,b)
Feature inheritance	Herzog and Koch (2001)	Hamker (2007)
Saccadic suppression of displacement	Deubel et al. (1996); Ziesche and Hamker (2014)	Ziesche and Hamker (2014)
Motion perception	Nishida (2011); Cavanagh (1992)	Bayerl and Neumann (2004, 2007b,a); Beck and Neumann (2010); Raudies and Neumann (2009, 2010); Raudies et al. (2011); Tsotsos et al. (2005)
Border ownership, texture boundaries, and object-based attention	Qiu et al. (2007); Voorhees and Poggio (1988); Roelfsema et al. (1998)	Thielscher and Neumann (2005, 2007); Thielscher et al. (2008); Thielscher and Neumann (2008); Poort et al. (2012); Weidenbacher and Neumann (2009)
Depth perception	Viswanathan and Mingolla (2002)	Bruce and Tsotsos (2012); Thielscher and Neumann (2008); Weidenbacher and Neumann (2009)

Table 2.1.: Visual attention in psychophysics and related models. The table lists psychophysical phenomena in which visual attention is involved, including a reference explaining the phenomenon in relation to attention. The last column lists all neuro-computational, system-level models of visual attention that can account for attention effects in a certain phenomenon, or explain the whole phenomenon via attention.

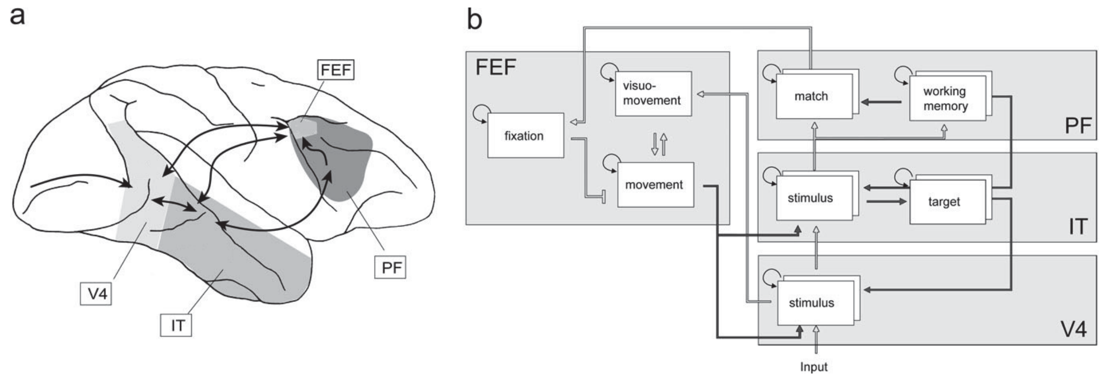


Figure 2.7.: The idea of emergent attention models. Such models propose that attention emerges from the neuronal dynamics between multiple brain areas due to the recurrent processing between them, and due to the amplification and suppression of neuronal activity. For example, the model of Hamker (2005a,b) simulates neuronal dynamics between high-level areas of the ventral stream (V4, IT), the frontal eye field (FEF), and the prefrontal cortex (PF). The illustration in **a)** shows the brain areas, while **b)** shows the model. Reprinted from Hamker (2005a).

These models can be classified primary in two major branches (Tsotsos and Rothenstein, 2011) according to their underlying general principles: emergent attention models and selective routing models. From the Tsotsos and Rothenstein (2011) classification, we leave out saliency models as they were already discussed in Sec. 2.4, and temporal tagging as they contain relatively aged models and temporal effects are already listed in Sec. 2.3.

The emergent attention models propose that attention emerges as a result of neuronal dynamics between multiple brain areas (Fig. 2.7). They typically include ventral (e.g. V4, IT) or dorsal stream areas (e.g. MT), spatial related areas like LIP or FEF, and prefrontal cortex areas. The models simulate the amplifying and suppressing effects of attention on the neuronal activity, which leads in combination with recurrent connections to an emergence of attention. The models belong mainly to four research groups: Hamker, Neumann, Lanyon, and Deco. Each group has a slightly different research direction as outlined in the following paragraphs. Some other research groups have also developed system-level models that are capable of attentional processing, but focus on the modeling of the visual cortex instead of attentional processing: adaptive resonance theory from Carpenter and Grossberg (ART, Carpenter et al. (2005); Grossberg (2013)), and VisNet (Rolls, 2012). The proposed model will belong as well to the class of emergent attention models.

Hamker's group focuses on spatial emergent of attention in combination with eye movements, and on guided visual search to localize objects. For example, spatial attention is deployed shortly before a saccade to the saccade target location (selective attention in visual search, Hamker (2005a,b)), amplifying there neuronal responses which compresses the space (spatial compression, Hamker et al. (2008)). Furthermore, receptive fields are transferred before the saccade to the future position on the retina (predictive remapping of receptive fields, Ziesche and Hamker (2014)), and are shifted towards the saccade target (shift of receptive fields, Zirnsak et al. (2011b)). Guided visual search describes the phenomenon to search for a given stimulus, which is the typical task to search, to locate, and to recognize a target object in a scene (Wolfe (1994) or Sec. 2.4). In the phenomenon, feature-based attention is deployed towards the target, which amplifies target-related and suppresses distractor-related activity. This modulated activity in the visual cortex provides a target specific bias to the ocular-motor system, resulting in a saccade to the target object (Hamker, 2005a). Hamker's group publishes their models in the following list of articles: Antonelli et al. (2014); Beuth et al. (2010, 2014); Fix et al. (2010); Hamker (2004b,a, 2005a,b,c, 2006); Hamker and Zirnsak (2006); Hamker (2007); Hamker et al. (2008, 2011); Ziesche and Hamker (2011, 2014); Zirnsak et al. (2010); Zirnsak and Hamker (2010); Zirnsak et al. (2011a,b).

Neumann's group focuses on motion perception or texture boundaries. In motion perception, V1 neurons react for local motion patters while the higher area MT estimates from this activity the global motion of a stimulus. However, a moving stimulus will also evoke responses in V1 neurons encoding an incorrect motion direction or speed. Attention helps to suppress these wrong responses (motion disambiguating, (Bayerl and Neumann, 2004)). Area MT estimates the motion direction and speed encoded by the majority of V1 neurons, and feeds this information back to V1 that amplifies neurons encoding the correct motion properties and suppresses incorrect ones. This recurrent processing leads to a clear neuronal code encoding the correct motion of a stimulus. Attention within a similar recurrent processing helps to detect boundaries (Thielscher and Neumann, 2007). V2 respond to local boundaries segments and a higher area (e.g. V4) forms a consistent boundary representation from this. This neuronal activation projects a feedback signal backwards, leading to an amplification of correct boundary segments and to a suppression of incorrect ones. After detecting the boundaries, feedback is deployed to neurons within the object, which amplifies the texture of the object and any related properties (region filling-in, Poort et al. (2012)). This is similar to the effect of object-based attention (Roelfsema et al., 1998): attention spreads over an object defined by luminance contrast (Egly et al., 1994), color (Mitchell et al., 2003; Reynolds and Desimone, 2003), or illusory contours (Moore

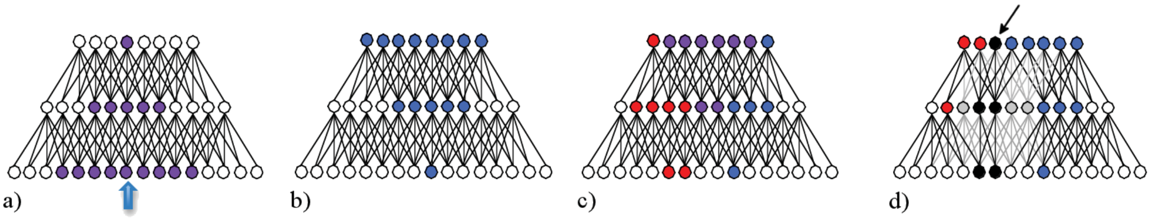


Figure 2.8.: The selective routing hypothesis addresses the signal routing problems in the cortex between low and high-level cortical areas. **a)** The context problem: A single neuron in the high-level area receives input from many neurons in the low-level areas, thus it is unknown if it is driven by the attended stimulus (arrow) or by an adjacent, contextual one. **b)** The blurring problem: The activity evoked by a single stimulus in a single low-level neuron drives many high level neurons, resulting in a loss of localization information. **c)** The cross-talk problem: Two sensory events result in a non-overlapping neuronal activity in low-level areas, but will lead to an overlapping neuronal activity in the high-level areas. **d)** According to the hypothesis, attention solves these problems by routing information selectively through the cortex. This is here shown in the scenario of c) when a top-level neuron, marked with the arrow, is attended (or selected). Reprinted from Tsotsos and Rothenstein (2011).

et al., 1998). In summary, the following system-level models belong the the group of Neumann: Bayerl and Neumann (2004, 2007b,a); Beck and Neumann (2010); Neumann and Sepp (1999); Raudies and Neumann (2010); Raudies et al. (2011); Poort et al. (2012); Thielscher and Neumann (2005, 2007); Thielscher et al. (2008); Thielscher and Neumann (2008); Weidenbacher and Neumann (2009).

Lanyon investigates visual attention in only two publications: Lanyon and Denham (2004) focus on guided visual search in a similar way as in Hamker (2005b), while Lanyon and Denham (2009) center on the temporal dynamics of spatial attention in biased competition in area IT (Chelazzi et al., 1998).

Deco focuses in his early works on visual attention (Corchs and Deco, 2002; Deco et al., 2002; Deco and Rolls, 2004), while later he moves on to executive attention (Deco and Rolls, 2005). In his early works, he has been developed for example an especially large model with all major ventral stream areas (V1, V2, V4, IT; Corchs and Deco (2002)), which explains the biased competition data of Reynolds et al. (1999) and Kastner et al. (1999). He also extended the VisNet framework of Rolls (Wallis and Rolls, 1997; Rolls, 2012) with top-down attentional processing, and simulated with this model biased competition, visual search, and the increasing magnitude of attentional modulation from low- to high-level areas (Deco and Rolls, 2004).

The second big branch of models centers around the selective routing hypothesis (Tsotsos et al., 1995; Tsotsos and Rothenstein, 2011), which addresses the questions how sensory information is routed from low- to high-level cortical areas, and how task-related information is projected from high- to low-level areas. It sees attention as a mechanism to solve the arising issues: 1) A single neuron in a high-level area receives input from many neurons in low-level areas, thus it is unknown if it is driven by the attended stimulus or by an adjacent, contextual one (context problem, Fig. 2.8a). 2) The activity evoked by a single stimulus in a single low-level neuron drives many high level neurons, resulting in a loss of localization information (blurring problem, Fig. 2.8b). 3) Two sensory events result in a non-overlapping neuronal activity in low-level areas, but will lead to an overlapping neuronal activity in the high-level areas (cross-talk problem, Fig. 2.8c). Attention solves these issues by routing information selectively through the cortex (Fig. 2.8d). The hypothesis is primarily developed by the research group of Tsotsos via their selective tuning models: Bruce and Tsotsos (2012); Rodríguez-Sánchez et al. (2007); Rothenstein et al. (2008); Rothenstein and Tsotsos (2014); Tsotsos et al. (1995, 2005, 2008); Zaharescu (2004).

A selecting tuning model (Tsotsos et al., 1995) consists of a hierarchy of areas, whereby each area contains layerwise-ordered interpretive, gating, and bias units (Rothenstein and Tsotsos, 2014). Interpretive units encode features and form so a feedforward network. This is complemented with a feedback network of gating units, which execute hierarchical winner-take-all processes (WTA processes). Each WTA process inspects the synaptic inputs of an interpretive neuron, gates the ones contributing at most to the response, and disables all others. After the process has finished in the top layer, the gating information is projected back to the next lower layer, starting WTA processes there, and so on. At the end, the gating network routes only the most significant signals through the model, i.e. the stimulus information which determines the responses in the top layer. The gating is task-unspecific, instead task information is provided via bias units, forming a second feedback network. They multiplicatively suppress activity of interpretive neurons and so bias them for the current task.

Besides addressing the routing problem, the group of Tsotsos explains also psychophysical phenomena, primarily depth perception and visual search. They use feedback to solve the stereo correspondence problem during depth perception and to deploy attention in depth (Bruce and Tsotsos, 2012). V1 neurons detect local disparities (energy-model, Qian (1994); Read and Cumming (2006)), and a higher area creates from this a global disparity representation. This information is projected backwards to V1, leading to a gating of neurons with correct disparities and to a suppression of incorrect ones. Furthermore, the model segments the attended stimulus in the left and right image as the feedback is projected from V1 further back to the image layer, allowing to gate or to suppress single pixels. Tsotsos stated that the selective tuning model is rooted in neurophysiology, yet the link remains vague as his studies do not replicate the neurophysiology effects of attention. The recent work of Rothenstein and Tsotsos (2014) addresses this problem by elaborating a more concrete implementation of the selective tuning model at microcircuit level. They validated their work by simulating the temporal course of neuronal responses in biased competition (Reynolds et al., 1999). In summary, the strengths of the selecting tuning models reside in solving the routing problems and in powerful explanations of psychophysical phenomena. Yet, their neurophysiology foundation seems still weak as biased competition is the only data simulated by such a model.

To conclude, system-level models of visual attention are able to explain quite a lot of psychophysical phenomena (Tab. 2.1). However, some phenomena have been never simulated by system-level models, providing prospering, future research direction: object substitution and meta-contrast masking, crowding, attentional blink to some degree, and multiple object tracking. The missing modeling attempt in object substitution masking is one of our motivations to simulate this phenomenon with a system-level model of attention (Chap. 5).

2.6. Conclusion

In this chapter, we gave an overview of the different branches of visual attention models. We review: 1) single-area models explaining neurophysiological data, 2) by attention inspired computer vision systems for object localization, and 3) neuro-computational system-level models explaining psychophysics and general principles of attentional processing (Sec. 2.3 - Sec. 2.5). A few of the neuro-computational models are also capable of object localization, thus there is some common ground between neuro-computational and computer vision models (Sec. 2.4.3). Aside from this, the model branches are completely divergent, although they all address the same topic of visual attention. We would like to solve this dilemma with our novel attention approach as it will explain neurophysiology, computer vision, and psychophysics with a common set of neuronal mechanisms.

3. A mechanistic cortical microcircuit of attention for amplification, normalization and suppression

Abstract Computational models of visual attention have replicated a large number of data from visual attention experiments. However, typically each computational model has been shown to account for only a few data sets. We developed a novel model of attention, particularly focused on explaining single cell recordings in multiple brain areas, to better understand the underlying computational circuits of attention involved in: spatial- and feature-based biased competition, modulation of the contrast response function, modulation of the neuronal tuning curve, and modulation of surround suppression. In contrast to previous models, we use a two layer structure inspired by the layered cortical architecture. We found out that four neuronal mechanisms are essential for attention: 1) amplification of the response of a neuron preferring an attended stimulus, 2) divisive normalization of the response, 3) spatial pooling within the receptive field of a neuron, and 4) suppression from neurons encoding different features or different locations in the surround.

This chapter in full has been published in “Vision Research” as the following article: Beuth and Hamker (2015a).

3.1. Introduction

Attention is one of the fields in vision research that has been strongly influenced by models. While early models have been rather abstract, over the last 15 years, inspired by data from neural recordings, the neural correlates of attention have been increasingly addressed by neuro-computational models. Although attention operates across multiple brain areas and may help to bind stimulus properties processed in different parts of the brain, one important line of research in visual attention aims at understanding the local circuits of neural interactions.

A very influential milestone has been the discovery of interactions between representations of stimuli that are placed within a common receptive field - coined as biased competition (Desimone and Duncan, 1995; Desimone, 1998). According to this concept, attention should not be understood as a simple mechanism for gating or an enhancement of neural responses, but rather as a competition for neural representation which is typically strong if neurons in a single area have spatially overlapping receptive fields. Thus, if attention is directed away, the neural response to a single stimulus is typically larger than the response to two different stimuli, which suggests that the total stimulus energy is not additive. Attention towards a particular location in space or towards a stimulus feature implements a top-down signal which changes the properties of local interactive circuits almost as if the attended stimulus would be presented alone, i.e. the response of neurons tuned to the attended stimulus is amplified and the response of neurons tuned to the not attended one is reduced. Neural recordings in different brain areas have supported this basic concept (Chelazzi et al., 1998, 2001; Desimone and Duncan, 1995; Fallah et al., 2007; Motter, 1993; Lee and Maunsell, 2009, 2010a,b; Reynolds et al., 1999; Treue and Trujillo, 1999; Zhou and Desimone, 2011). Other observations of the neural effect of attention revealed multiplicative scaling (McAdams and Maunsell, 1999; David et al., 2008; Motter, 1993) or sharpening (David et al., 2008; Martinez-Trujillo and Treue, 2004) effects on the neural tuning curve. Moreover, several studies observed that attention typically leads to a shift in the contrast response function (Reynolds et al., 2000; Martínez-Trujillo and Treue, 2002; Li et al., 2008) which suggests that well visible stimuli are not further boosted. A large number of neuro-computational models have been developed and demonstrated to account for parts of these data (Ardid et al., 2007; Boynton, 2009; Buia and Tiesinga, 2008; Compte and Wang, 2006; Hamker, 2004b, 2005a; Hugues and José, 2010; Lee and Maunsell, 2009; Ni et al., 2012; Reynolds and Heeger, 2009; Spratling and Johnson, 2004; Spratling, 2008; Wagatsuma et al., 2013). Basically, each of these models includes lateral or feedforward inhibition and some form of attentive gain increase. Further experimental studies shed more light on the nature of stimulus interactions, e.g. by probing the neural responses to two identical stimuli (Lee and Maunsell, 2009, 2010a) or led to conflicting observations of whether attention shifts or scales the contrast response function (Martínez-Trujillo and Treue, 2002), which cumulated into normalization models of attention (Lee and Maunsell, 2009; Reynolds and Heeger, 2009). Although these models are not completely different than their predecessors, they propose two essential components that determine the final population response, an attentive drive that scales the response multiplicatively and a suppressive drive that operates after the attentive drive within a local region and acts divisively on the neural response. The combination of these two components allowed to account for modeling competitive interactions of two identical stimuli (Lee and Maunsell, 2009, 2010a) and to solve the apparent conflict whether attention shifts or scales the contrast response function (Reynolds and Heeger, 2009).

Although Reynolds and Heeger (2009) demonstrated that their model can account for several data, existing models of attention have so far shown to account only for a very selective set of data. However, a particular strength of computational models is their ability to reveal a few potential underlying computational mechanisms that may account for a large variety of data. Thus, from the computational point of view it is important to find out how much data can be explained by a small set of computational mechanisms. The mechanisms of attention can be explained by different levels of implementation detail, e.g. detailed biophysical neuron models and their microcircuits, abstract spiking neurons, dynamic rate coded populations or more abstract mathematical descriptions. With respect to the available data we here developed a cortical microcircuit of attention using dynamic rate coded neural populations inspired by properties of the cortical microcircuit (Douglas and Martin, 2004). This leaves the exact biophysical implementation unanswered but allows to reveal the main computational mechanisms that may similarly operate in multiple brain areas regardless of their exact implementation. Although we mainly address data from neuronal recordings, the chosen level of implementation detail allows to discuss and better understand the influence of attention in psychophysical experiments, such as the effect of spatial and feature-based attention, target selection, and distractor inhibition.

3.2. Methods

The mechanistic cortical microcircuit of attention is an attempt to account for a very large dataset and thus to reveal the essential aspects of attention common in multiple brain areas. We identify three core mechanisms necessary for this approach: amplification, normalization, and suppression. A set of Matlab routines implementing the model can be downloaded at: `https://www.tu-chemnitz.de/cs/KI/supplement/Beuth2019`.

3.2.1. Model overview and structure

The model (Fig. 3.1a) is structured in a modulating (cortical layer 4) and in a pool (cortical layer 2/3) layer. A neuron in this model abstracts from a biophysical description of a cell

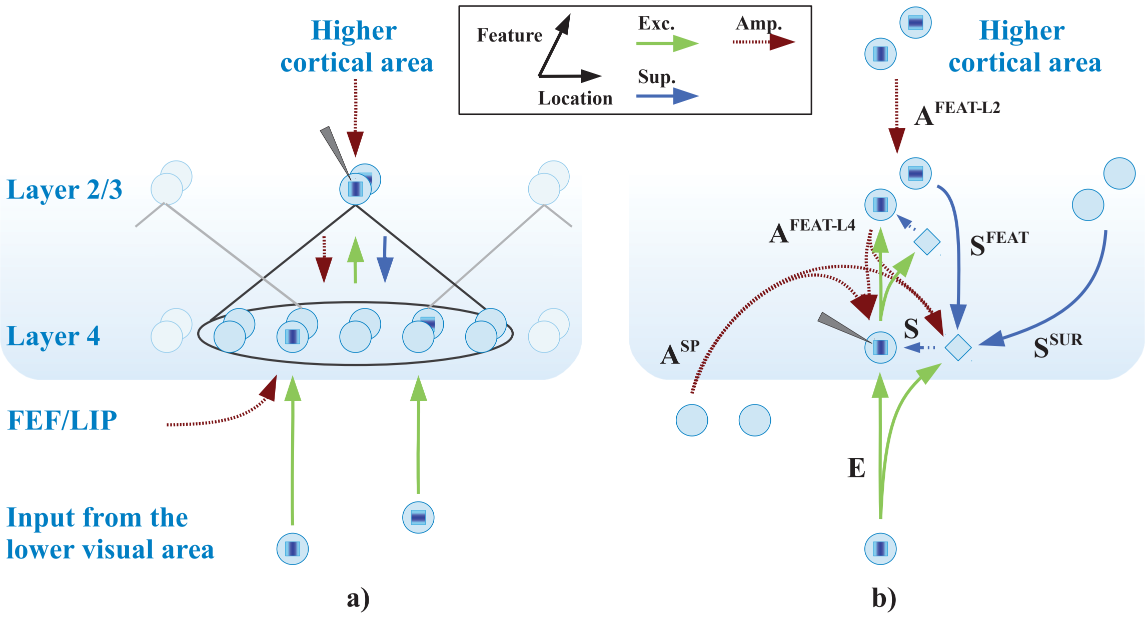


Figure 3.1.: **a)** Proposed mechanistic cortical microcircuit of attention, consisting of a modulatory layer 4 and a spatial pooling layer 2/3. Feature-based attention can be deployed to layer 2/3 via an amplifying signal originating from a higher cortical area, e.g. PFC or IT. Spatial attention can be deployed to layer 4 via an amplifying signal originating from cortices encoding spatial information, e.g. FEF or LIP. **b)** Connectivity and influences centered on a single cell in layer 4, indicated by the electrode symbol. The cell receives feedforward excitation (E), feature-based amplification ($A^{\text{FEAT-L4}}$) from layer 2/3, spatial amplification (A^{SP}) from spatial cortices such as FEF or LIP, and suppression from an associated interneuron (S). The interneuron (symbol $\diamond$) receives several sources of suppression: the feedforward excitation of its associated neuron (E); dissimilar features in layer 2/3 (S^{FEAT}); and all similar features in the surround in layer 2/3 (S^{SUR}).

and rather refers to a population of interconnected real cells. Similarly, gain amplifications are described by multiplicative operations. This level of description is similar to proposed concepts of normalization (Carandini and Heeger, 2012) and the normalization model of attention (Reynolds and Heeger, 2009), although we will suggest in more detail how abstract concepts of normalization and suppression may be implemented in a simple cortical microcircuit of attention. A neuron in each layer encodes a certain feature at a particular (one-dimensional) location, so that a layer of neurons is represented by a two dimensional matrix. The mechanistic cortical microcircuit of attention has three inputs: layer 4 receives input from lower visual areas, from areas that encode space such as the frontal eye field (FEF) and the lateral intraparietal cortex (LIP), and layer 2/3 receives a signal from a feature selective higher cortical area. The model mainly refers to physiological properties of area V4 and MT. However, as attention is a canonical property of the visual cortex (Saenz et al., 2002), the model is not restricted to these areas and may account also for other visual areas.

The tuning curve of a neuron has to be determined by the user to match the particular characteristics of a brain area as tuning curves largely differ across visual areas. Layer 4 is denoted as modulating layer as its neuronal responses are calculated by amplifying or suppressing the excitation given from the input. Neurons in layer 4 converge onto neurons in layer 2/3, which spatially pool over them in a broad area (a similar idea as in Hamker and Zirnsak (2006); Compte and Wang (2006)). Therefore, layer 2/3 neurons have larger receptive fields than layer 4 neurons. The receptive fields in layer 2/3 are overlapping by a user defined amount (standard is 50% to each side), so a stimulus will fall within several receptive fields. Dependent on the amount of overlap and the receptive field sizes, layer 2/3 contains typically less cells in the spatial dimension than layer 4. In all experiments, we report the responses of cells in layer 2/3 (also indicated by the electrode symbol in each figure).

3.2.2. Divisive normalization

The neural responses in the model largely follow concepts of divisive normalization (Carandini and Heeger, 2012; Reynolds and Heeger, 2009). As a concretization of the original concept used by Reynolds and Heeger (2009), we propose an implementation via interneurons. In this implementation, each pyramidal neuron receives via an inhibitory connection a normalization signal from an associated interneuron (Fig. 3.1b, S), which normalizes the response of the pyramidal neuron. The interneuron receives suppressive influences ($S^{\text{FEAT}}, S^{\text{SUR}}$) via excitatory connections plus the feedforward excitation of the associated pyramidal neuron. This is supported by the data from Mitchell et al. (2007) who found that attention modulates interneurons twice as strongly as pyramidal cells, and by the anatomical meta-analysis from Potjans and Diesmann (2012) reporting the required inhibitory, intra-layer connections from an interneuron to a pyramidal neuron, and the necessary excitatory connections from an pyramidal neuron to an interneuron or another pyramidal neuron.

3.2.3. Modeling of attention via amplification

Attention is thought to emerge from recurrent connectivity within the visual and prefrontal cortex (Hamker, 2005a). This computational framework proposes that neurons receive feedback from higher cortical areas, which acts as an amplifying signal. To distinguish between mechanism and result, we will use "attention" to describe the psychological effect and "amplification" to refer to a particular mechanism at the single neuron.

We assume that spatial attention implies a spatial amplification signal from spatial maps in the brain, such as the FEF (Hamker, 2005a) or the LIP (Steenrod et al., 2013), denoted in the model as A^{SP} (Fig. 3.1b). In the model, this signal amplifies the response of neurons in layer 4 (similar as in Hamker (2005a) and Spratling (2008)) as inspired by anatomical observations, e.g. layer 4 of area V4 receives input from the FEF (Barone et al., 2000) and LIP (Ninomiya et al., 2012). By this mechanism, spatial amplification can be directed to one or multiple spatial locations within the receptive field as needed in experiments with spatial attention.

We presume that feature-based attention implies a feature-specific amplification of neuronal responses in the visual cortex, typically invariant to spatial location. This amplification can occur via an amplifying signal originating from prefrontal cortex (PFC) which is then projected backwards through the visual cortex. This assumption is consistent with the attentional control of the visual cortex as reviewed by Kastner and Ungerleider (2000) for feature-based, and by Miller and Buschman (2013) for spatial and feature-based attention. Therefore, we model a feature-based amplification signal originating from a higher visual area, e.g. PFC or IT, which is projected to V4 or MT. This signal is firstly projected to layer 2/3 via excitatory connections (Fig. 3.1b, signal $A^{\text{FEAT-L2}}$) and then afterwards sent to layer 4 (Fig. 1b, signal $A^{\text{FEAT-L4}}$). We assume the projection from layer 2/3 to layer 4 occurs indirectly via the cortical excitatory pathways of layer 2/3 to layer 5 and then to layer 4. Both projections were found in the physiological connectivity data of the microcircuit (Douglas and Martin, 2004) which appears to generalize across areas in the visual cortex. In layer 4, the effects of both spatial and feature-based amplification signals are summed up additively because this behavior was observed in several areas in the visual cortex (Treue and Trujillo (1999): area MT; Saenz et al. (2002): areas V1, V2, V3A, V4, and MST).

3.2.4. Modeling of suppression

Suppression from other neurons (Fig. 3.1b, signals S^{FEAT} and S^{SUR}) is modeled via connections from layer 2/3 to layer 4. To our knowledge, it is not known in detail which intra-area connections are responsible for transporting suppressive signals, however layer 2/3 is beneficial as starting point because feedback connections from higher areas end in layer 2/3 and feedback seems to play an essential role in suppression (Angelucci et al., 2002; Bair et al., 2003; Carandini and Heeger, 2012; Gilbert, 1998; Hunt et al., 2011). Layer 4 is favorable as end point as its neurons need to receive suppression to balance out the amplification. To keep the model simple, suppression is implemented by two separate signals, each representing a particular experimental condition: feature-based suppression within the classical receptive field (S^{FEAT}), and surround suppression from outside the receptive field (S^{SUR}). This facilitates the usability of the model as the amount of suppression can be fitted to a particular experiment without influencing the other condition.

3.2.5. Mathematical description of the model

Notation

Firing rates of mean-rate neurons are denoted by r: r^{In} (Input), r^{L4} (layer 4), r^{L2} (layer 2/3), r^{PFC} (higher cortical area like PFC or IT), and r^{FEF} (spatial areas like FEF or LIP). The rates r^{In}, r^{PFC}, and r^{FEF} have to be set by the user to model a particular experimental condition. Big letters denote influences towards a neuron: E denotes the excitatory influence, A the amplifying one, and S the suppressive one. Each neuron is addressed within a layer by the spatial index x and the feature index l. The layer dimensions are denoted by X and L respectively. The feature dimension is modeled circular, hence the features $l = 1$ and $l = L$ encode adjacent features. As the model is a recurrent system, we simulate it via ordinary differential equations over time, denoted by t. All firing rates were recorded in a steady state condition after the system has converged, which typically occurs after 100ms.

Influences towards a neuron and connectivity

A neuron is influenced via several connections, which are designed either as one-to-one connections or as complex connectivity patterns. A one-to-one connection connects a single presynaptic cell with a single postsynaptic cell. Both cells encode the same location and feature. Influences via one-to-one connections:

Excitation $E^{\text{In-L4}}$ (input $\rightarrow$ layer 4) Excitatory connection from input layer to layer 4 with the connection strength $v^{\text{In-L4}}$.

Amplification A^{SP} (FEF/LIP $\rightarrow$ layer 4) Amplifying connection from a spatial cortical area, e.g. FEF or LIP, to layer 4 with the strength $v^{\text{FEF-L4}}$.

Amplification $A^{\text{FEAT-L2}}$ (higher cortical area $\rightarrow$ layer 2) Amplifying connection from a higher cortical area, like PFC or IT, to layer 2/3 with the strength $v^{\text{PFC-L2}}$.

A complex connectivity pattern connects multiple presynaptic cells to a single postsynaptic cell, thus multiple presynaptic features or locations are integrated to a single postsynaptic feature and location. We modeled such an influence by an equation like $E_{post} = v \cdot \sum_{pre}(w_{post,pre} \cdot r_{pre})$, using a global scaling factor v and a connectivity matrix $w_{post,pre} \in [0,1]$. This modeling via two separate variables was chosen as the experimental data can be well fitted by adapting only the scaling factor v and keeping the connectivity pattern w unchanged. Influences via complex connectivity patterns:

Excitation $E^{\text{L4-L2}}$ (layer 4 $\rightarrow$ layer 2/3) The connectivity pattern $w^{\text{L4-L2}}_{x,x'}$ for each postsynaptic cell (x, l) is a Gaussian over the spatial dimension of the receptive field (x') with constant feature (l). The parameters of the Gaussian are chosen so that the Gaussian is centered and that it has a value of about 0.5 at the borders of the receptive field: $\mu = 0$, $\sigma = 0.4 \cdot s$, whereby $s = 5$ denotes the spatial extend of the field. Connections exist only inside the receptive field, except the experiment of Cavanaugh et al. (2002a) which needs a broader connectivity pattern $(2\,s)$. The scaling factor of the excitation is fixed to 1.

Amplification $A^{\text{FEAT-L4}}$ (layer 2/3 $\rightarrow$ layer 4) The connectivity pattern $w^{\text{L2-L4}}_{x,x'}$ for each postsynaptic cell (x, l) is a Gaussian over the spatial dimension of the receptive field (x') with constant feature (l). It is reciprocal modeled to $w^{\text{L4-L2}}$. Scaling factor: $v^{\text{L2-L4}}$.

Feature-based suppression S^{FEAT} (layer 2/3 $\rightarrow$ layer 4) The connectivity pattern w^{FEAT}_{dl} from a presynaptic cell with feature l' at location x to a postsynaptic cell with feature l is a linear function (Eq. 3.1). The function is 0 if pre- and postsynaptic features are equal $(dl \leq L_0)$, and 1 if they are at most dissimilar. Thereby, dl denotes the distance in feature space. As the feature dimension is modeled circular, the features are most dissimilar at medial distance ($dl = L/2$) in the feature space $[1, L]$. Scaling factor: v^{FEAT}.

$$w_{dl}^{\text{FEAT}} = \begin{cases} 0 & dl \leq L_0 \\ \frac{dl - L_0}{0.5L - L_0} & L_0 < dl \leq 0.5\,L \\ 1 - \frac{dl - 0.5L}{0.5L - L_0} & 0.5\,L < dl \leq L - L_0 \\ 0 & else \end{cases} \tag{3.1}$$

whereby $dl = |l - l'|$ and $L_0 = L/8$.

Surround suppression S^{SUR} (layer 2/3 → layer 4) The connectivity pattern $w_{x,x'}^{\text{SUR}}$ for each postsynaptic cell (x, l) is a function (Eq. 3.2) over the spatial dimension of the presynaptic layer (x') with constant feature (l). It is 0 for the same location, maximal (1.0) for close surround locations and moderate (0.4) for distant ones. Scaling factor: v^{SUR}.

$$w_{x,x'}^{\text{SUR}} = \begin{cases} \frac{|x-x'|-1}{\psi - 1} & |x - x'| \leq \psi \\ 1 - 0.6\,\frac{|x-x'|-\psi-1}{\psi-1} & \psi < |x - x'| \leq 2\,\psi \\ 0.4 & else \end{cases} \tag{3.2}$$

whereby $\psi = (s - 1)/2$ denotes the radius of the receptive field s.

Neuronal layers

The neuronal responses are simulated by the following equations. The values of parameters varied across experiments are listed in Tab. 3.1, and the ones kept constant are the following: $\sigma^{\text{L4}} = 0.3$, $\sigma^{\text{L2}} = 2$, $v^{\text{L2-L4}} = 1$, $\tau = 10ms$.

Layer 4

$$\begin{aligned} \tau \frac{\partial r_{x,l}^{\text{L4}}}{\partial t} &= g \cdot \frac{E_{x,l} \cdot A_{x,l}}{\sigma^{\text{L4}} + S_{x,l}} - r_{x,l}^{\text{L4}} && (3.3) \\ S_{x,l} &= E_{x,l} \cdot A_{x,l} + b\,(E_{x,l}) \cdot S_{x,l}^{\text{SUR}} + b\,(E_{x,l}) \cdot S_{x,l}^{\text{FEAT}} && (3.4) \\ A_{x,l} &= 1 + A_{x,l}^{\text{SP}} + A_{x,l}^{\text{FEAT-L4}} && (3.5) \\ E_{x,l} &= \left(v^{\text{In-L4}} \cdot r_{x,l}^{\text{In}}\right)^{p_E} && (3.6) \\ \text{With: } A_{x,l}^{\text{SP}} &= v^{\text{FEF-L4}} \cdot r_{x}^{\text{FEF}} && (3.7) \end{aligned}$$

$$\text{With: } A^{\text{FEAT-L4}}_{x,l} = v^{\text{L2-L4}} \cdot f_1\left(\sum_{x'} f_2\left(w^{\text{L2-L4}}_{x,x'} \cdot r^{\text{L2}}_{x',l}\right)\right) \tag{3.8}$$

$$\text{With: } S^{\text{SUR}}_{x,l} = \sum_{x'} w^{\text{SUR}}_{x,x'} \cdot \left[v^{\text{SUR}} \cdot r^{\text{L2}}_{x',l}\right]^{p_{SUR}} \tag{3.9}$$

$$\text{With: } S^{\text{FEAT}}_{x,l} = \sum_{l'} w^{\text{FEAT}}_{l,l'} \cdot \left[v^{\text{FEAT}} \cdot f_1\left(\sum_{x'} f_2\left(w^{\text{L2-L4}}_{x,x'} \cdot r^{\text{L2}}_{x',l'}\right)\right)\right]^{p_{FEAT}} \tag{3.10}$$

$$\text{With: } b(E_{x,l}) = E_{x,l} \cdot \left(\frac{\sigma^{\text{L4}} + E_{x,l}}{g \cdot E_{x,l}}\right)^2 \tag{3.11}$$

Whereby $b(E_{x,l})$ represents a correction function based on the excitation ($E_{x,l}$) and the inverted divisive normalization function without modulation, i.e without amplification and suppression (terms in brackets). The function ensures that suppression causes the same proportional decrease of the response for all stimulus contrasts. The parameters σ^{L4}, $v^{\text{In-L4}}$ and p^{E} control the shape of the contrast response function similar as in Carandini and Heeger (2012). The parameters p^{FEAT} and p^{SUR} control the non-linearity behavior of feature-based and surround suppression.

The factor g normalizes the firing rates to a maximum of 1 as in other divisive normalization approaches (Albrecht and Hamilton, 1982; Carandini and Heeger, 2012). It should be chosen as: $g = 1 + \sigma/(A \cdot E')$ whereby $A \cdot E'$ denotes the maximum possible excitation (here 1).

Layer 2/3

$$\tau \frac{\partial r^{\text{L2}}_{x,l}}{\partial t} = g \cdot \frac{E_{x,l} \cdot \left(1 + A^{\text{FEAT-L2}}_{x,l}\right)}{\sigma^{\text{L2}} + S_{x,l}} - r^{\text{L2}}_{x,l} \tag{3.12}$$

$$S_{x,l} = E_{x,l} \cdot \left(1 + A^{\text{FEAT-L2}}_{x,l}\right) \tag{3.13}$$

$$\text{With: } E_{x,l} = f_1\left(\sum_{x'} f_2\left(w^{\text{L4-L2}}_{x,x'} \cdot r^{\text{L4}}_{x',l}\right)\right) \tag{3.14}$$

$$\text{With: } A^{\text{FEAT-L2}}_{x,l} = v^{\text{PFC-L2}} \cdot r^{\text{PFC}}_{l} \tag{3.15}$$

$$\text{With: } f_1(x) = \frac{p^{\text{Pool}}}{4} \cdot x^{\frac{1}{p^{\text{Pool}}}} \text{ and } f_2(x) = x^{p_{Pool}} \tag{3.16}$$

Whereby $f_1(x)$ and $f_2(x)$ represent non-linearities allowing the non-linear summation necessary for a spatial pooling operation. The pooling operation is modeled after the pooling

Fig.	E			S^{FEAT}		S^{SUR}		A		β	Tuning function
no.	$v^{\text{In-L4}}$	p^{E}	p^{Pool}	v^{FEAT}	p^{FEAT}	v^{SUR}	p^{SUR}	$v^{\text{FEF-L4}}$	$v^{\text{PFC-L2}}$		
Std.	3	2	4	3	2	0.5	1	3	0.5	0	$c = \frac{8}{L}, a = 0$
3.2				2.5				2		0.15	$c = \frac{6}{L}$
3.3a				2.75				1			$c = \frac{6}{L}$
3.3b				3.5							$c = \frac{8}{L}, a = 0.15$
3.3c-e				3.5	3			1.25			$c = \frac{8}{L}, a = 0.1$
3.3f	3.5			2.5						0.05	$c = \frac{8}{L}, a = 0.15$
3.4					3				0.17	0.15	$c = \frac{8}{L}, a = 0.1$
3.5	2	1.75									$c = \frac{8}{L}$
3.6						1					$c = \frac{8}{L}$
3.7	2	1.5				0.75				0.1	$c = \frac{8}{L}, a = 0.1$
3.8a				2	3						$c = \frac{8}{L}, a = 0.3$, Amplification: $c = \frac{4}{L}$
3.8b				2.4	3.5				0.25		$c = \frac{6}{L}$
3.9a	0.5	0.75	2			0.8	2				$c = \frac{8}{L}$
3.9b	1.6	1.25				0.55	2.5				$c = \frac{8}{L}$
3.10											$c = \frac{8}{L}$

Table 3.1.: Model parameters that were varied between experiments are listed in the table. The second row "std" denotes the standard values that were used in all experiments if not otherwise denoted. The input is modeled via a symmetrical exponential function in the feature space with width c and baseline a: $f(x) = a + (1 - a) \cdot e^{(-c \cdot x)}$. If not otherwise noted, the amplification is modeled via the same parameterized function, but without the baseline ($a = 0$). The model contains also an unmodulated baseline activity β to simulate conditions in which attention does not modulate the response.

and surround suppression data of Bonin et al. (2005); Heuer and Britten (2002); Hunter and Born (2011); Jones et al. (2002); Pack et al. (2005). It is chosen as $p^{\text{Pool}} = 4$.

Baseline activity can be modulated by attention like in the sharpening of neuronal tuning curves (no. 3.8a in Tab. 3.1), but is also often unaffected by attention. To model the former case, we add a small value to the input (parameter a). For the latter case, we add a small value (parameter β) to the firing rates after the differential equation system has been converged (Eq. 3.17, 3.18). Our approach is similar as in the normalization model of attention (Reynolds and Heeger, 2009) to facilitate a comparison of the data fits.

$$r^{\text{L4}} = \beta + (1 - \beta) \cdot r^{\text{L4}} \tag{3.17}$$

$$r^{\text{L2}} = \beta + (1 - \beta) \cdot r^{\text{L2}} \tag{3.18}$$

3.3. Results

In the following, we will identify essential mechanisms of our model for each attention experiment. We see these mechanisms as a proposal of how attention might work in the cortical microcircuit of the visual cortex. In each experiment, we first verify the model by comparing its simulated neuronal responses to physiological data and then link the simulated data to the mechanisms in the model. Further, we will predict attention mechanisms in novel paradigms, e.g. in biased competition with surround suppression.

3.3.1. Biased competition

In biased competition paradigms, suppression is observed between two different stimuli presented inside a receptive field of a neuron and attention has been observed to modulate this suppression. Biased competition with spatial attention has been reported in several areas (MT: Lee and Maunsell (2010a), also Lee and Maunsell (2009, 2010b); V2 and V4: Reynolds et al. (1999); V1, V2, V4: Motter (1993)). In Lee and Maunsell (2010a), two inversely moving gratings were placed in the receptive field of a neuron in area MT (Fig. 3.2a). Attention was either spatially directed to a grating placed in the opposite hemifield (attend away) or to one of the two stimuli. They observed that an unattended stimulus alone elicits a much stronger neuronal response in its preferred neuron (column 1 left) as the unattended stimulus pair (column 3). However, if the preferred stimulus of the pair (column 4) is attended, the response of the preferred cell will be as strong as to the stimulus alone (column 4, left) and the response of the anti-preferred cell will be decreased (column 4, right). The reverse effect is observable for attending the anti-preferred stimulus (column 5). The data is adapted from Lee and Maunsell (2010a, Fig. 5), and shows the average firing rate over the full population and over a time window between 50ms and 200ms after stimulus onset.

In the model (Fig. 3.2b), spatial attention is modeled by an amplification signal originating from FEF/LIP to layer 4, amplifying the response of all cells at the attended location. This location comprises only the attended stimulus, here the upward moving one, so only the response of this feature is increased. This drives layer 2/3 cells, resulting also there in an amplified response of this feature as observed in the data (column 4, left in Fig. 3.2a). These neurons in layer 2/3 suppress cells in layer 4 encoding the anti-preferred stimulus (downward motion), denoted as feature-based suppression. This in turn decreases also the

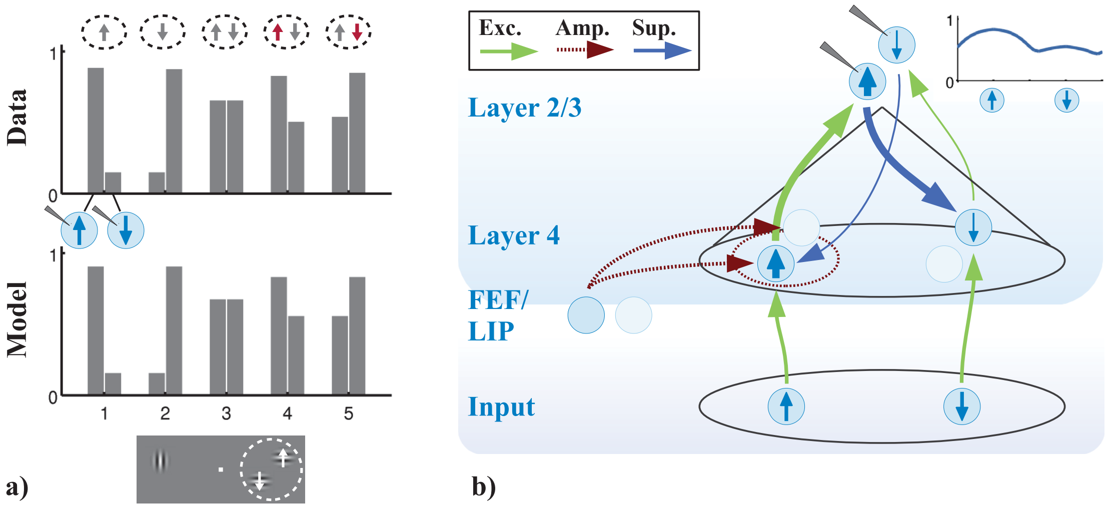


Figure 3.2.: Biased competition experiment with spatial attention. **a)** Neurophysiological data (top) in relation to our simulation results (bottom). Data is adapted from Lee and Maunsell (2010a, Fig5). In each column, the left bar denotes the response of a neuron preferring the upward moving stimulus and the right bar the downward direction. Spatial attention is either directed away (columns 1 - 3) or to the location of the stimulus marked in red (columns 4 and 5). **b)** Model mechanism of biased competition, within the condition of spatially attending the left location (condition 4 in a). In each layer, exemplary neurons reacting to one of the two stimuli are shown. In addition, the full population response of layer 2/3 at the recorded location is illustrated at the top right. The connection type (excitation, amplification, suppression; see Sec. 3.2) describes the influence on its postsynaptic neuron. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness. For clarity, interneurons and unrelated connections are not illustrated in the figure.

response of cells in layer 2/3 encoding the anti-preferred stimulus as observed in model and data (column 4, right in Fig. 3.2a).

For simulating biased competition with spatial attention, the spatial pooling operation in the second layer is necessary. This operation is one of the main model improvements as previous models contain only a single layer (Boynton, 2009; Lee and Maunsell, 2009; Ni et al., 2012; Reynolds and Heeger, 2009) as discussed in more detail later. Pooling mediates the interaction of the attended location with the recorded location. The recorded location is here defined as the center of the receptive field, because the model uses symmetric receptive fields. The attended location is classically in the periphery of the receptive field. Therefore, spatial pooling is required to transfer the attentional modulated response from the attended location to the center. In addition, spatial pooling is accompanied by a distribution of the suppression. This is implemented via the suppressive feedback connections from layer 2/3 to layer 4 originating from the center to all locations within the receptive field. Both mechanisms together allow that a spatial amplification at one location induces a suppression at the other location.

Therefore, necessary mechanisms for biased competition are the amplification in layer 4 and in layer 2/3, the suppression of the anti-preferred feature, and the spatial pooling.

Biased competition has been traditionally tested with different stimuli raising the question what effect occurs if identical stimuli are placed within the receptive field. Recent data suggests that a "feature pooling" effect occurs instead of biased competition. The effect without attention is shown by MacEvoy et al. (2009) for area V1: no suppression was observed between equal stimuli, i.e. the response to a single stimulus was the same as presenting both stimuli (Fig. 3.3b, right). In the study, they varied systematically the similarity between the stimuli and showed that the amount of suppression decreases with increasing similarity (Fig. 3.3b). Yet it has to be noted that they used overlapping stimuli whereby the stimuli in biased competition are normally spatially distinct. Nevertheless, the pooling effect also occurs with deploying attention to one of two spatially distinct stimuli in area MT: Lee and Maunsell (2010a) reported for two equal stimuli a similar attentional modulation as for a single stimulus (Fig. 3.3a right). The effect could not be induced by the experimental setup as the stimulus similarity has been systematically varied within the same setup, and biased competition has been found for different stimuli and pooling for identical stimuli. Our model reflects this effect as the strength of the suppressive connections from layer 2/3 to layer 4 depends on the feature similarity (Fig. 3.1), so the strength is maximal for different ones and zero for similar features.

A possibility to quantitatively measure biased competition is by relating the response of a cell to a single stimulus (selectivity) to its response to a stimulus pair (sensory interaction, Reynolds et al. (1999)). The selectivity of a neuron is defined as the response to a variable probe stimulus minus the response of an arbitrarily chosen reference stimulus. A positive selectivity refers to cells preferring the probe, while a negative value refers to cells preferring the reference stimulus. The sensory interaction is defined as the response to the stimulus pair (probe and reference presented simultaneously) minus the reference alone. A positive value indicates that the cell's response to the reference is increased by adding a probe, whereas a negative value denotes suppression by adding the probe. The selectivity and sensory interaction indices for several randomly chosen probe and reference stimuli have been plotted cell-wise against each other and fitted with a linear regression

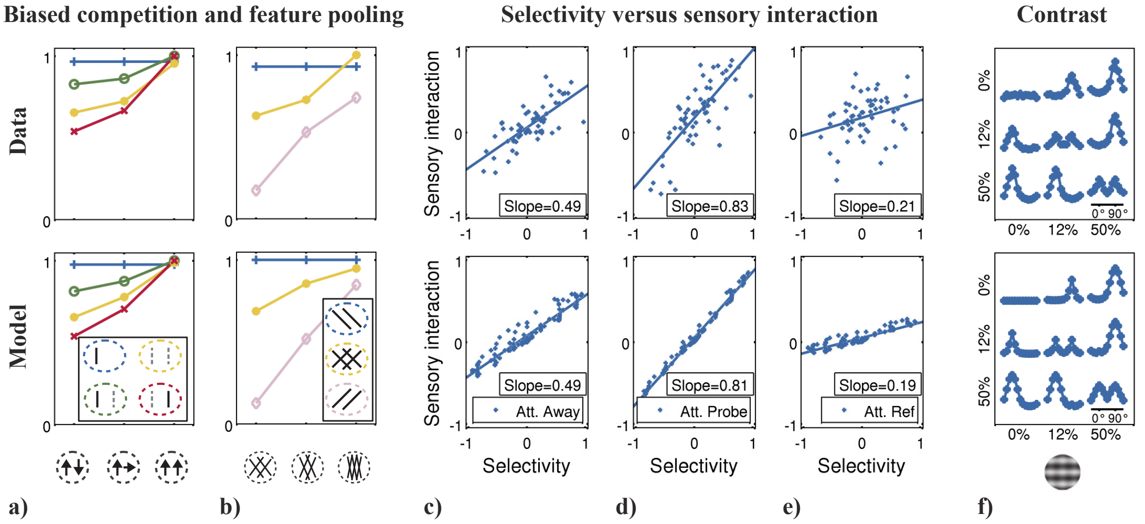


Figure 3.3.: Characteristics of biased competition and related paradigms. Neurophysiological data (top row) is shown in relation to simulation results (bottom row). **a)** Neuronal response depends on the similarity of the two spatially disjunctive stimuli. Biased competition (two different stimuli, left) and feature pooling (two equal stimuli, right) mark the extreme cases. Experimental conditions are: preferred stimulus alone attended (blue), pair attend preferred (green), pair attend away (yellow), and pair attend anti-preferred (red). The data shows the average population response in area MT of cells preferring the left stimulus and is adapted from Lee and Maunsell (2010a, Fig. 5). **b)** The same similarity dependency as in a) can be observed also for two overlapping gratings in area V1. In this study (MacEvoy et al., 2009, Fig. 2a-c), conditions are all without attention: preferred stimulus alone (blue), pair (yellow), and anti-preferred stimulus alone (pink). **c-e)** Effect of attention on the relationship between selectivity (probe - reference) and sensory-interaction (pair - reference). See text for details. The data represents a population response in area V4 in a biased competition paradigm with spatial attention (Reynolds et al., 1999, Fig. 11a,b,d). **f)** Contrast dependency of responses in area V1 with overlapping competing stimuli (Busse et al., 2009, Fig. 4a). The contrast of two stimuli (0° and 90° oriented gratings) is systematically varied (0%, 12%, 50%) and the population response is plotted. Contrasts of horizontal gratings (90°) are denoted at the x-axis and of vertical gratings (0°) at the y-axis.

curve. In area V4, the slope in the unattended condition is 0.49 (Fig. 3.3c). A slope of 1.0 would indicate no suppression between the stimuli, hence the slope of 0.49 shows for a cell preferring the probe (positive selectivity) a suppression from the reference of about 50%. Likewise for a cell preferring the reference (negative selectivity), it shows a suppression from the probe of about 50%. Attending the probe results in a slope of 0.83 (Fig. 3.3d), which demonstrates an increase to the pair response for cells preferring the probe (positive selectivity) and a decrease to the pair response for cells preferring the reference (negative selectivity). Contrary, attending the reference (Fig. 3.3e) will evoke the opposite effect and thus resulting in a slope of 0.21. Additionally to area V4, Reynolds et al. (1999) investigated also area V2 and found a qualitatively similar effect. The model can replicate the data well (Fig. 3.3c-e) with the previously outlined mechanisms involved in biased competition.

All the above experiments were carried out with stimuli of equal contrast. However, the amount of suppression depends on the contrast of each stimulus. Stimulus contrast is defined in this context as the contrast between the stimulus and the background, hence it describes the strength of a stimulus. Busse et al. (2009) used overlapping gratings of 0° and 90° orientation, varied systematically the contrast of both orientations, and recorded the neuronal population via electrode arrays (Fig. 3.3f). If a single stimulus is presented alone (shown as 0% contrast of the other stimulus in Fig. 3.3f), the neuronal response is described by the contrast response function (CRF, Albrecht and Hamilton (1982)). The CRF is implemented in the model by divisive normalization in layer 4: $r^{\text{L4}} = (1 + \frac{1}{\sigma}) \cdot \frac{E}{\sigma+E}$, whereby the excitation E represents the stimulus contrast. This divisive normalization explains the strong non-linear influence of the contrast on the response. If a pair of differently contrasted stimuli is presented (12% and 50% contrast in Fig. 3.3f), the population response is dominated by the higher contrasted stimulus. This implies a non-linear suppression towards cells preferring the lower contrasted stimulus, realized in the model by the same feature-based suppression mechanism as in biased competition: suppression occurs from cells in layer 2/3 preferring the 50% contrasted stimulus towards cells in layer 4 preferring the 12% contrasted stimuli. Afterwards, the decreased response in layer 4 is projected to the recorded neuron in layer 2/3. The observed non-linear effect occurs because the suppression originates in a non-linear manner from the neuronal response in layer 2/3 (r^{L2}): $S = (r^{\text{L2}})^{p_{FEAT}}$; and because this suppression operates divisively on the firing rate in layer 4: $r^{\text{L4}} = (1 + \frac{1}{\sigma}) \cdot \frac{E}{\sigma+E+b(E)\cdot S}$. These mechanisms are present in the model in all conditions, but the non-linear effect is mostly visible in cases with different contrasted stimuli. With two equally contrasted stimuli, the data shows another effect, an equally strong suppression towards both stimuli and under all contrasts. This expected result proves that suppression occurs at all contrasts, which implies that also biased competition should occur at all contrasts.

Feature-based attention (review: Maunsell and Treue (2006)) can also modulate the suppression between stimuli in a biased competition paradigm, similar to the effect described

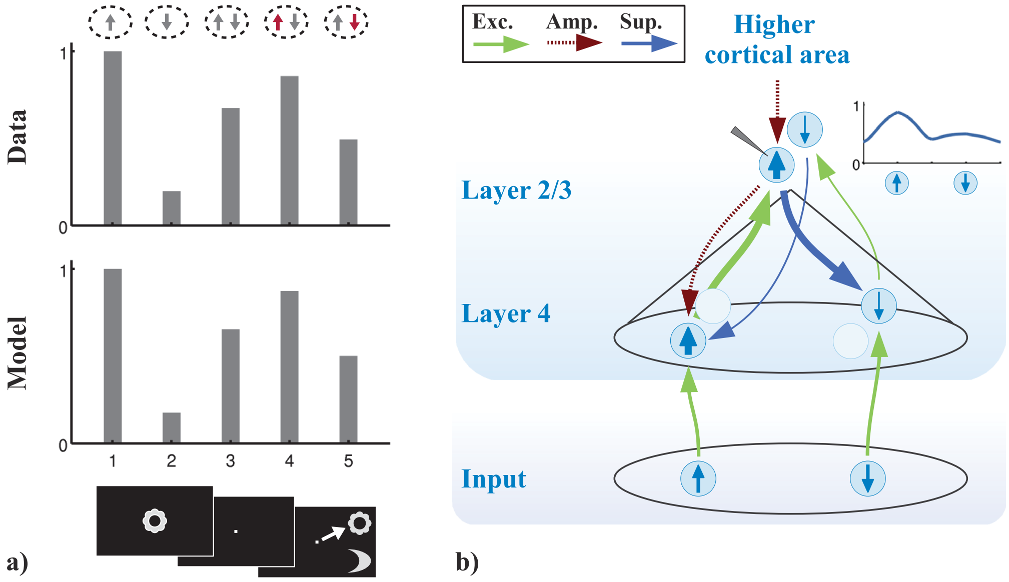


Figure 3.4.: Biased competition experiment with feature-based attention. **a)** Neurophysiological data (top) in relation to our simulation results (bottom, recorded cell is indicated in (b)). The average neuronal response of area V4 for the preferred stimulus (1), anti-preferred stimulus (2), and pair (3-5). In the latter, feature-based attention (denoted by red) is either directed away (3), to the preferred (4), or to the anti-preferred stimulus (5). Data adapted from Chelazzi et al. (1998, Fig. 10a and 10b). **b)** Mechanics of biased competition with feature-based attention. The notation of the figure is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness. The mechanics are based on the experiment illustrated in (a), condition 4.

for spatial attention but with a feature-selective bias. Studies reported the effect for area V4 (Chelazzi et al., 2001; Fallah et al., 2007; Zhou and Desimone, 2011), MT (Treue and Trujillo, 1999), and IT (Chelazzi et al., 1998). In the study of Chelazzi et al. (1998), the task of three rhesus monkeys was to execute a saccade to a previously shown target object that was presented together with distracting objects (Fig. 3.4a). The target object was previously shown at a different location to ensure that the setup causes only feature-based attention. The presented data (Fig. 3.4a, top) illustrates the population response of neurons preferring the target in area IT. The response was obtained by averaging over 58 neurons and over a time span from 150ms to 300ms after stimulus onset. The time window was chosen so that it starts at the onset of the full response and ends before the neuronal response changes due to onset of the saccade, similar as used by Lee and Maunsell (2010a) for their setup of biased competition. The data demonstrates the typical biased competition effects: the unattended, preferred stimulus alone (Fig. 3.4a, condition 1) elicits a much stronger neuronal response than the unattended stimulus pair (condition 3). However, if the preferred stimulus of the pair is attended, the response is amplified (condition 4), whereas a decrease of the response is observed if the anti-preferred stimulus is attended (condition 5). Yet, it is still higher as the response to the unattended, anti-preferred stimulus alone (condition 2). Although this task initially deploys feature-based attention as the location of the target object is varied across trials, it has been proposed and modeled that saccade preparation induces a delayed spatial reentrant signal that affects the late response (Hamker, 2005a). However, as we here focus on the cortical microcircuit of attention, we simplify the modeling and only apply a feature-specific bias.

The mechanisms of biased competition with feature-based attention are illustrated in Fig. 3.4b. The search array composed of target and distractor excites respective cells in layer 4 and layer 2/3. A feature-based amplification originates from a higher cortical area, and is projected to layer 2/3 amplifying the neuronal response of the attended feature. This in turn enhances via amplifying connections the neuronal response of this feature in layer 4, and decreases via suppressive connections the response of the distracting feature in layer 4, which also reduces the excitatory signal to layer 2/3 neurons selective for the distracting feature.

3.3.2. Attentional modulation of the contrast response function: contrast gain or response gain

The amount of attentional modulation depends on the stimulus contrast. Contrast gain (definition: Reynolds et al. (2000)) describes that attention amplifies the neuronal response mainly at lower stimulus contrasts, whereas at high contrast, the neuronal response saturates to the same value without an explicit threshold on the response. This results in a leftward shift of the contrast response function (CRF, Albrecht and Hamilton (1982)). The response gain denotes the appealing simple idea that attention amplifies the neuronal response by a fixed factor, as already suggested very early by McAdams and Maunsell (1999) or Treue and Trujillo (1999). There has been experimental evidence for both, contrast gain and response gain, as explained in the following.

Reynolds et al. (2000) observed contrast gain in single-cell recordings of area V4 involving spatial attention. Their data (Fig. 3.5a) illustrates the contrast response function of a

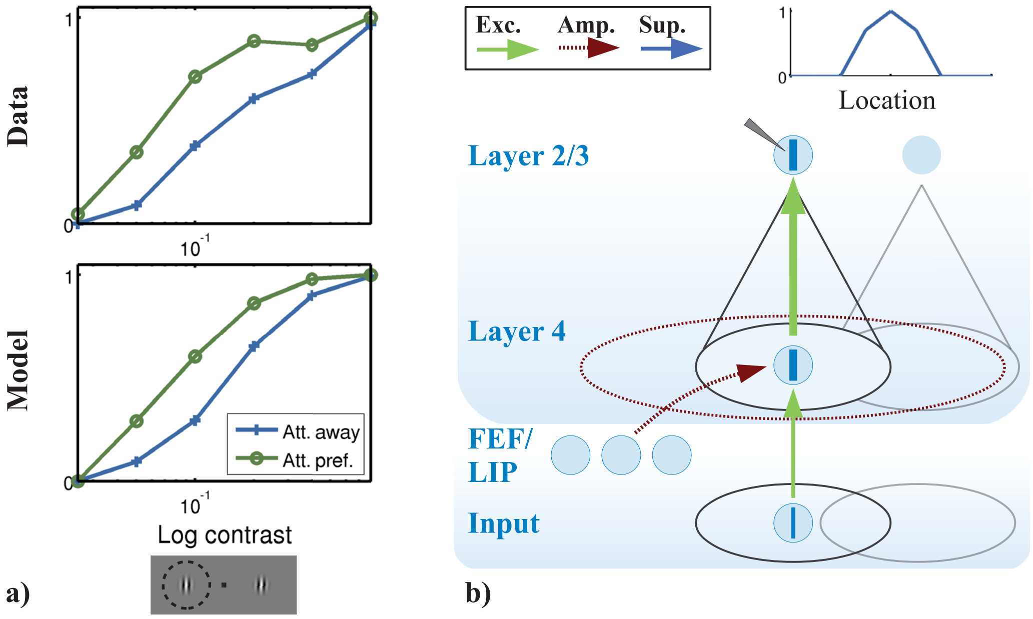


Figure 3.5.: Contrast gain resulting from the attentional modulation of the contrast function. **a)** Neurophysiological data (top) in relation to simulation results (bottom, recorded cell is indicated in (b)). The average neuronal response of area V4 as a function of contrast (data adapted from Reynolds et al. (2000, Fig. 5a). **b)** Mechanics of contrast gain. Neuronal activity in layer 2/3 occurs only at a small region as illustrated by the spatial response profile at the top right, showing across space the response of neurons preferring the stimulus. The notation of the figure is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

population of V4 neurons to a grating. In this study, the task of two rhesus monkeys was to spatially attend inside (attend preferred, Fig. 3.5a left) or outside the receptive field of the recorded cell (attend away, Fig. 3.5a right).

Reynolds and Heeger (2009) already suggested by means of model simulations that whether attention leads to contrast gain or response gain depends on the stimulus size in relation to the attended region. Reynolds et al. (2000) used a rectangle as a cue that was much larger than the stimulus, presumably resulting in a broader field of attention compared to the size of the stimulus. Besides area V4, Li et al. (2008) reported contrast gain in area V1 with fMRI using spatial attention. Furthermore, Martínez-Trujillo and Treue (2002) reported the effect in area MT for exemplary neurons, but found also neurons with a response gain.

In our model, contrast gain occurs if the attended region is much larger than the stimulus (Fig. 3.5b). Thus, spatial attention is deployed very broadly to layer 4, but amplifies

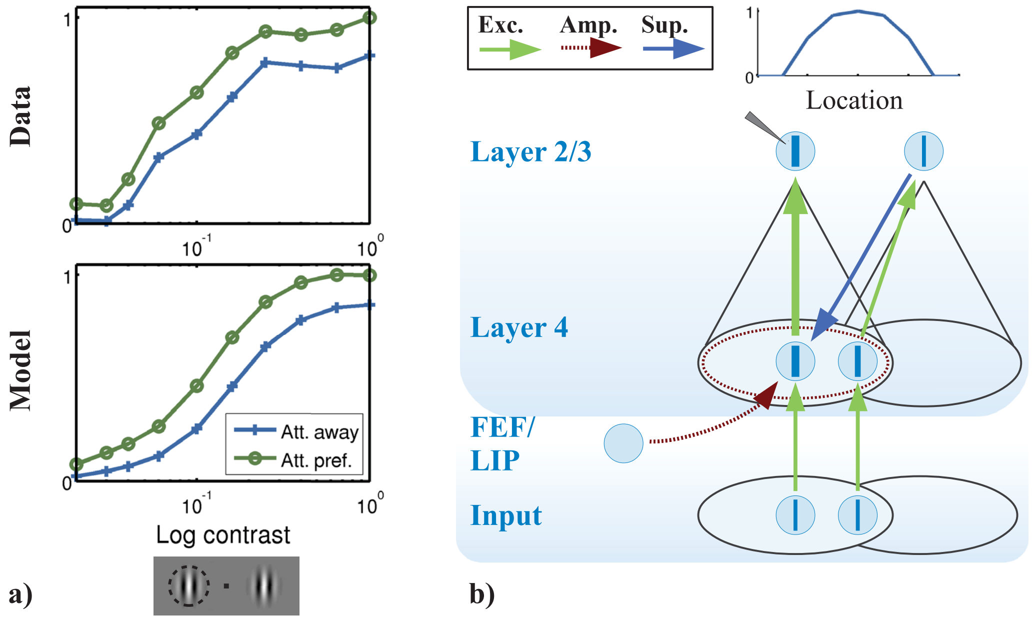


Figure 3.6.: A mix of response and contrast gain resulting from the attentional modulation of the contrast response function. **a)** Neurophysiological data (top) in relation to simulation results (bottom, recorded cell is indicated in b). The average neuronal response of area V4 as a function of contrast (CRF, data adapted from Williford and Maunsell (2006, Fig. 6g)). **b)** Mechanisms responsible for a mix of response and contrast gain. Neuronal activity in layer 2/3 occurs at multiple locations as illustrated by the spatial response profile at the top right, showing across space the response of neurons preferring the stimulus. The notation of the figure is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

only a few neurons as only a few features at a single location will be excited due to the small stimulus size. This response is projected to layer 2/3, activating only neurons within a very narrow region. Thus no suppression from neurons at spatially adjacent locations occurs, allowing the response to saturate. At high contrast, the response saturates to its maximal possible value, so attentional amplification occurs only at low contrasts. Both effects together, the saturation at high and the attentional amplification at low contrasts, induces the typical leftwards shift of the contrast response function (Fig. 3.5a).

A mixed effect between contrast and response gain was observed by Williford and Maunsell (2006), in area V4. Furthermore, evidence for response gain has been found in fMRI data by Boynton (2009) in areas V1, V2; by Li et al. (2008) in areas V2, V3, V3A, V4s; and by Murray (2008) in areas V1, V2, V3. Williford and Maunsell (2006) recorded the contrast response function (CRF, Albrecht and Hamilton (1982)) for a population of V4 neurons, and obtained an average CRF (Fig. 3.6a) over all neurons with a significant attentional modulation and a significant contrast enhancement of the CRF. The task of two macaque monkeys was to spatially attend to a grating either inside the receptive field of the recorded cell (attend preferred, Fig. 3.6a left stimulus) or outside of it (attend away, Fig. 3.6a right). The cueing stimulus was a grating of the same size as the stimuli, presumably resulting in an attended area that has roughly the same size as the stimulus.

In our model, spatial attention leads to a pure response gain if the stimulus is much broader than the attended area (similar as in Reynolds and Heeger (2009)), and to a mix of response and contrast gain if the stimulus and the attended area are similar in size (Fig. 3.6b). In both cases, the stimulus falls within the receptive fields of multiple, spatially-adjacent neurons in layer 2/3. In the unattended case, the broad responses in layer 2/3 result in a suppression to layer 4 which in turn results in a moderate response in layer 4 (left cell) and also in layer 2/3 (left cell). This occurs similarly at all contrast levels. The suppression occurs between different locations similar as surround suppression, thus we model it via the same signal A^{SUR}. In the attended case, attention is deployed to only a small area amplifying only there the neuronal response. Thus, the previously moderate response is amplified. As this mechanism is contrast independent, the response gain is observed under all contrasts.

A response gain is observed in many of the other simulated experiments (biased competition: Fig. 3.2, 3.4; attentional modulation of neuronal tuning curve: Fig. 3.7, 3.8; attentional modulation of surround suppression: Fig. 3.9b). In some experiments, it is shown directly by the modulation of the contrast response function (Fig. 3.9b). In the other experiments, only high contrasted stimuli were presented. In this case, an increase of the neuronal response through attention indicates the response gain effect (Fig. 3.2, 3.4, 3.7, 3.8). According to our model, a response gain is observed if suppression is similar strong among different contrasts, as observed for example in the study of Lee and Maunsell (2010b) (same data as Lee and Maunsell (2010a) shown in Fig. 3.2). The setup involves spatial attention and the neuronal responses in the conditions attend preferred and attend anti-preferred were recorded. A response gain is observed because this setup induces similar effects across all contrasts: In the condition attend anti-preferred, a strong suppression from layer 2/3 to layer 4 results in a moderate response (Fig. 3.2); and in the attend preferred condition, the neuronal responses layer 4 are amplified. As amplification and suppression mechanisms modulate the responses similarly strong for all contrasts, a response gain effect is observed.

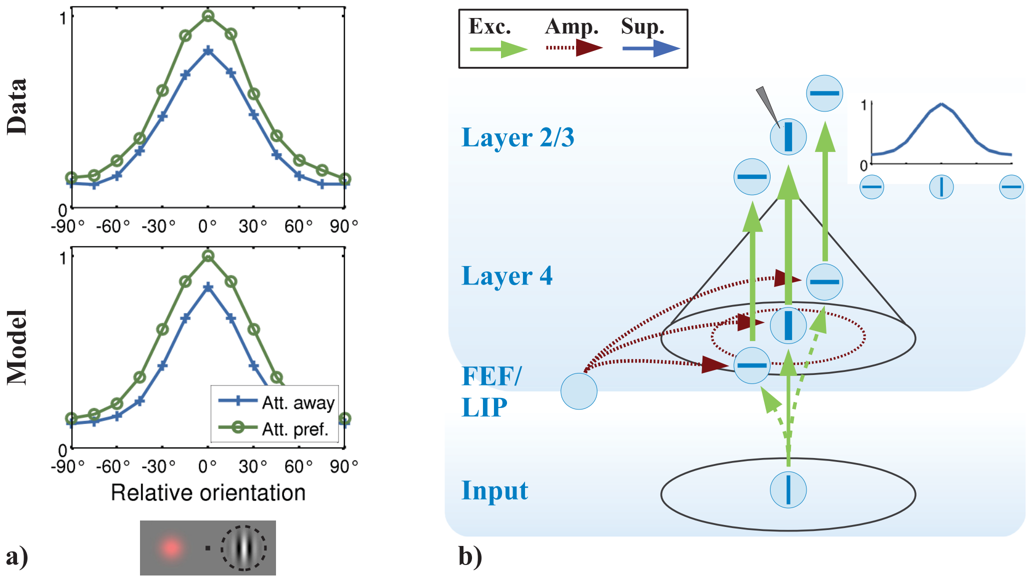


Figure 3.7.: Scaling of the neuronal tuning curve due to spatial attention. **a)** Neurophysiological data (top) in relation to simulation results (bottom, recorded cell is indicated in b). Color denotes conditions of attend away (blue) and attend preferred (green). Data adapted from McAdams and Maunsell (1999, Fig. 4). **b)** Mechanics: spatial amplification enhances all features at the attended location with a similar factor, resulting in a multiplicative scaling of the tuning curve. The notation of the figure is similar to Fig. 3.2. The variable stimulus excites each neuron differently strong according to its feature preference (indicated by dotted arrows). Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

3.3.3. Attention modulation of neuronal tuning curves

Scaling of the neuronal tuning curve

Attention also affects the neuronal tuning curve in the feature space, resulting in either a scaling (Fig. 3.7a) or sharpening (Fig. 3.8a) of the curve (Ling et al., 2009).

Typically, the deployment of spatial attention results in a multiplicative scaling of the tuning curve, as observed by single-unit recordings in V1 (Motter, 1993), V2 (Motter, 1993), and V4 (David et al., 2008; McAdams and Maunsell, 1999; Motter, 1993). We simulated the study of McAdams and Maunsell (1999) who have recorded tuning curves in area V4 (Fig. 3.7a). In their study, two rhesus monkeys have to either report the color equality/difference of successive blob-stimuli presented in the left hemifield (condition: attend away) or the equality/difference of successive presented gratings presented in the right hemifield (condition: attend preferred). The stimulus in the right hemifield was placed within the receptive field of the recorded neuron. Furthermore, its orientation was varied to obtain the neuronal tuning curve. The data shows the average population response, which was acquired by shifting the preferred tuning of each single neuron to the center of the plot and then afterwards averaging the responses.

According to our model, the scaling effect occurs because spatial attention amplifies the response of all features within a population in layer 4 with a similar factor (Fig. 3.7b). The increased activity is projected to layer 2/3 neurons. Feature-based suppression is not important in this paradigm as the amount of suppression is relatively similar in both experimental conditions attend away and attend preferred. The amount of suppression depends on the differences between the neuronal responses within layer 2/3, because a cell receives feature-based amplification from its own feature and suppression from opposing ones. As these response differences do not change with spatial attention, the suppression remains constant between both conditions. In addition, suppression from adjacent neurons occurs as the stimulus is large enough to fall within several receptive fields, similar as in the response gain experiment (Fig. 3.6). In summary, suppression occurs, but the effect is primarily based on the very crucial property of attention, the amplification mechanism.

Sharpening of the neuronal tuning curve

Feature-based attention leads to an amplification of the attended feature and to a suppression of unattended ones, known as sharpening (Fig. 3.8a), as suggested by single-unit recordings in MT (Martinez-Trujillo and Treue, 2004) and V4 (David et al., 2008). Our simulations suggest that the effect is a general property of attention in the cortical microcircuit as it is based on two very basic properties: the amplification mechanism and suppression.

Martinez-Trujillo and Treue (2004) illustrate this effect for a representative neuron in area MT (Fig. 3.8a). Their experimental setup contains two identically moving random dot patterns (RDPs) at two locations (right location was recorded). In the task, two macaque monkeys have either to spatially attend to the fixation point (condition: attend away; Fig. 3.8a blue) or to the moving RDP at location 'left' (condition: attend preferred; Fig. 3.8a green). This setup ensures feature-based attention at the 'right' position: spatially attending the left RDP induces spatial and feature-based attention to it, presumably resulting in

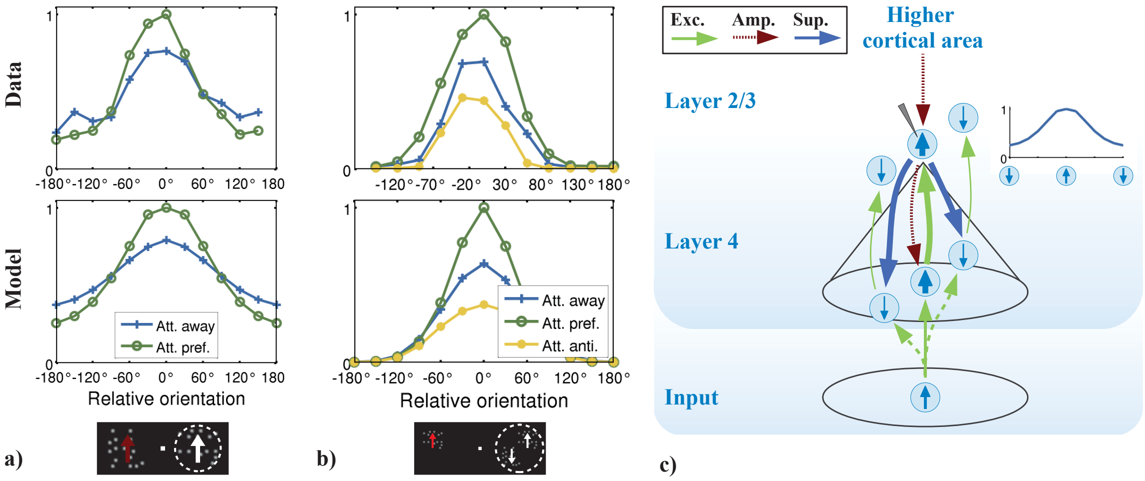


Figure 3.8.: Modulation of neuronal tuning curves due to feature-based attention. **a-b)** Neurophysiological data (top) in relation to simulation results (bottom). **a)** Sharpening of the neuronal tuning curve due to feature-based attention in a single-stimulus setup. The recorded cell is indicated in c). The data illustrates the response of a representative MT neuron for different orientated stimuli (adapted from Martinez-Trujillo and Treue (2004, Fig. 4a). **b)** In biased competition, feature-based attention leads to a scaling of the neuronal tuning curve. Data is adapted from Treue and Trujillo (1999, Fig. 3b). **c)** Mechanisms: in the single-stimulus setup of (a), feature-based attention results in a sharpening as it amplifies the response of only the attended feature and increases suppression to opposing ones. The notation of the figure is similar to Fig. 3.2. The variable stimulus excites each neuron differently strong according to its feature preference (indicated by dotted arrows). Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

a deployment of feature-based attention to all identical RDP in the scene, hence also to the right RDP. The tuning curve is sharpened by attention as the neuronal response to the preferred motion direction is increased and the response to the anti-preferred direction is decreased.

In our model (Fig. 3.8c), sharpening occurs if feature-based attention is narrowly tuned and thus amplifies only the response of cells in layer 2/3 and layer 4 preferring the attended feature. Amplifying this response leads to a stronger suppression from layer 2/3 to the other (and non-attended) features in layer 4 resulting in a decreased neuronal response. This in turn decreases also the response of opposing features in layer 2/3 (opposing motion direction are at 180° and -180°). These suppression effects induce the sharpening of the tuning curve (Fig. 3.8a).

Neuronal tuning curve in biased competition

Treue and Trujillo (1999) investigated the modulation of the tuning curve within an additional distractor by using a biased competition paradigm with feature-based attention. Unexpectedly, this setup results in a scaling of the tuning curve instead of a sharpening (Fig. 3.8b). They reported neuronal tuning curves of a representative neuron in area MT, and placed two stimuli in the receptive field of a cell: a target random dot pattern (RDP) with variable motion direction to record a tuning curve and a distractor RDP with fixed anti-preferred motion direction of the recorded cell. Attention was deployed either to the target, or to the distractor by cueing a third RDP placed in the contralateral hemifield that had the same feature as the attended object. As cue and attended object have different locations, this setup ensures pure feature-based attention. The task of the two macaques was to detect a motion change in the target stimulus. The data shows a multiplicative scaling of the tuning curve when attending the target (up-scaling) or distractor (down-scaling).

A pure multiplicative scaling is surprising as other data shows a sharpening of the tuning curve (Fig. 3.8a), and as other models (Reynolds and Heeger, 2009) could not provide a solid explanation for this discrepancy. However, our modeling of the experiment clarifies that the sharpening effect does not occur due to a strong suppression from the additional distractor. The presence of the distractor suppresses the flanks of the tuning curve as the neuronal responses are zero for anti-preferred targets in the attended-away condition (Fig. 3.8b, blue curve at motion directions in the range: -180 to -120° and 120° to 180°). Contrary, these responses are non-zero in the data showing the sharpening (Fig. 3.8a). In the model, suppression is most powerful at the flanks of the tuning curve. It is weaker at the center due to competition between neurons preferring the target and the distractor. Due to this suppression of the flanks, a further sharpening does not occur.

3.3.4. Attentional modulation of surround suppression

Attention can also modulate the amount of received suppression in center surround experiments. Within a classical receptive field (cRF), attentional modulated suppression is well explained by the biased competition paradigm (Sec. 3.3.1). Thus, the surround paradigm is especially interesting as it addresses the question of attentionally modulated suppression beyond the cRF. In such center surround experiments (Gilbert, 1998), a stimulus in the surround of the cRF suppresses the cell's response as demonstrated by Cavanaugh et al.

(2002a) for area V1. They systematically increased the size of a grating beyond the size of the cRF, and observed that the average neuronal response starts to decrease after the stimulus size exceeds the cRF showing the suppression effect (Fig. 3.9a). Furthermore, they found a larger cRF at lower contrasts. The cRF size was defined in their study as the stimulus size invoking the maximum response (grating summation field). Our attention model can account for this effect if we presume a weaker surround suppression at lower contrasts by setting $p^{\text{SUR}} = 2.0$. At higher contrasts, the strong suppression removes layer 4 activity at the flanks of the stimulus area, leading to a narrow spatial response curve. Contrary at lower contrasts, the response remains broad in the spatial dimension. Layer 2/3 neurons integrate so over a broader activity at lower contrast, resulting in higher responses. This approach is similar to the "gain" model of Cavanaugh et al. (2002a) which fitted at best the data.

We have simulated the experiment as follow. The input was simulated via a Gaussian in the spatial dimension. Its standard deviation represents the area of the stimulus: $\sigma \approx \frac{\pi d^2}{4}$, with stimulus diameter d. Layer 2/3 neurons need to integrate information beyond their cRF borders, so layer 4 and layer 2/3 are weakly connected outside the cRF too. In all other experiments, the layers have for simplicity only connections inside a cRF. In both cases, the width of the cRF is the same as the connectivity weights are modeled via the same Gaussian.

Surround suppression is widely observed in the visual cortex: in LGN (Bonin et al., 2005); V1 (Cavanaugh et al., 2002a,b; Bair et al., 2003); V2 (Ito and Gilbert, 1999; Willmore et al., 2010); V4 (Sundberg et al., 2009); MT (Allman et al., 1985; Pack et al., 2005; Hunter and Born, 2011); MST (Orban, 2008); and LIP (Falkner et al., 2010). In IT, no study reported this effect, probably because receptive fields are very large, thus it is experimentally very difficult to place a stimulus outside the receptive field. However, Miller et al. (1993) found suppression between pair of stimuli within a receptive field.

The surround effect can also be modulated by attention as attending the center of the receptive field or the surround stimulus amplifies or respectively suppresses the neuronal response. Sundberg et al. (2009) observed this effect by recording neuronal responses in area V4 (Fig. 3.9b). Their setup (yellow, Fig. 3.9b) consists of three gratings with variable contrast, one in the opposite hemifield (left stimulus in Fig. 3.9b), one in the classical receptive field (cRF) of the recorded cell (right stimulus marked with a ring), and one in the remote surround of the cRF (lower right stimulus). Stimulus contrast was varied for the stimulus inside the cRF and was fixed to the maximum for the stimulus in the surround. As

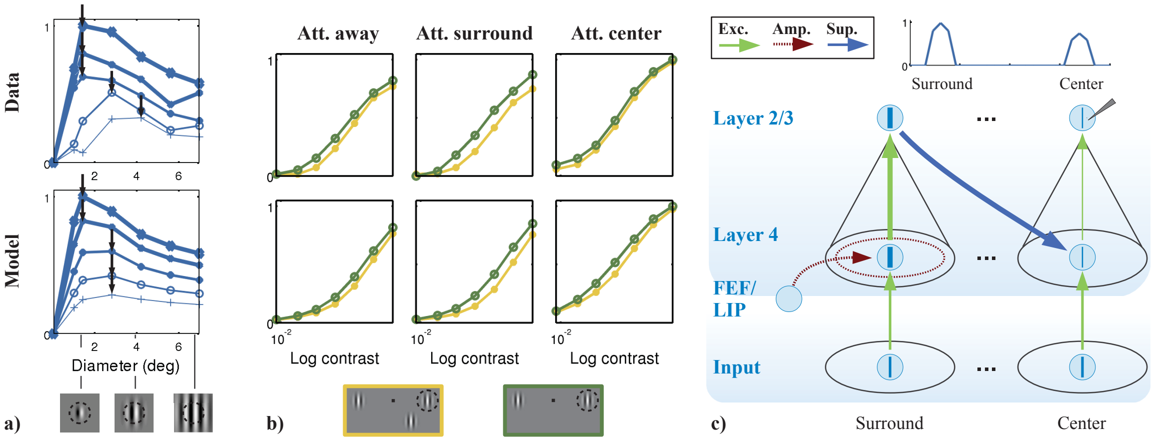


Figure 3.9.: Attentional modulation of surround suppression. **a - b)** Neurophysiological data (top) in relation to simulation results (bottom, recorded cells are indicated in c). a) The average neuronal response in area V1 for gratings of increasing diameter size (data adapted from Cavanaugh et al. (2002a, Fig. 8a). The stimulus contrast was also varied whereby the highest is denoted by the darkest points. The peak in a response curve marks the border of the receptive field, denoted by black arrows. **b)** The average response of neurons in area V4 as a function of contrast if one stimulus is presented in the center and another one in the surround of the classical receptive field (cRF). Yellow denotes this case (condition: surround), green denotes a control setup without the surround stimulus (condition: center). Data was adapted from Sundberg et al. (2009, Fig. 6f). **c)** Mechanisms of attentionally modulated surround suppression. The notation of the figure is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

a control (green), they replicate the known attentional modulation of the cRF for a single stimulus (Fig. 3.5 in Sec. 3.3.2). In both setups, they investigated the conditions of attending the grating in the opposite hemifield (attend away), in the center of the cRF (attend center, right stimulus), and in the surround (attend surround). They found that in the attend away condition, the surround stimulus slightly suppresses the neuronal responses (yellow versus green curve). Compared to the data of Cavanaugh et al. (2002a) shown in Fig. 3.9a, the weak suppression may be explained by the fact that Cavanaugh et al. (2002a) covers the whole surround with a large surround stimulus whereas Sundberg et al. (2009) uses a single and small surround stimulus. Attending the surround slightly increases the amount of surround suppression. In contrast, attending to the center stimulus diminishes the surround suppression (overlapping of the yellow and green curve). In addition, attending to the center stimulus increases the absolute response, indicating a response gain effect in addition to the surround effect.

The mechanics in the model will be explained by the example of the study of Sundberg et al. (2009) (Fig. 3.9c). The center and the surround stimulus excite neurons in layer 4 and in turn in layer 2/3 at both the surround (left in Fig. 3.9c) and the center location (right). The layer 2/3 response from the surround is projected back via suppressive connections to layer 4 in the center, decreasing the neuronal response which in turn also decreases the layer 2/3 activations in the center. As this suppression does not occur if no surround stimulus is present, its influence is visible by comparing the responses with and without surround stimulus (yellow versus green in Fig. 3.9b). An opposite suppression effect occurs also from the center to the surround location (not shown). If the surround is spatially attended, the responses in both layer 4 and 2/3 are amplified which increases the suppression to layer 4 in the center. Attending the center invokes the opposite mechanism, resulting in a decreased response at the surround.

Besides the study of Sundberg et al. (2009), a few other studies investigated the effect of attention in center surround experiments and found similar results: Chen and Martinez-Conde (2008) found the effect in area V1 by spatially attending the center or the surround with a variable task difficulty. On the highest difficulty, they found an amplification of about 140% when attending the center. Ito and Gilbert (1999) shows the effect also for area V1 in the conditions of attending away, the center, or four surround locations simultaneously.

Until now, we have considered only the surround suppression between similar stimuli and in cases where the suppression is independent of the spatial arrangement of the surround. This is the most typical case (Zanos et al., 2011) and so implemented as standard in the model. However in general, the effect is more complex as reviewed by Gilbert (1998) or Spratling (2010) (the latter study particularly focused on V1). Bonin et al. (2005) found the condition that suppression occurs from all features in the surround, independently of their similarity. This was also observed for random dot patterns, which create a strong suppression independent from spatial arrangement. However in the paradigm denoted as contour linkage (or flanker effect, Gilbert (1998)), adjacent stimuli can even enhance each other if they form together a large contour. The model might also account for these effects if the amplifying and suppressive connections from layer 2/3 to layer 4 (A^{FEAT} and S^{SUR}) are adjusted appropriately.

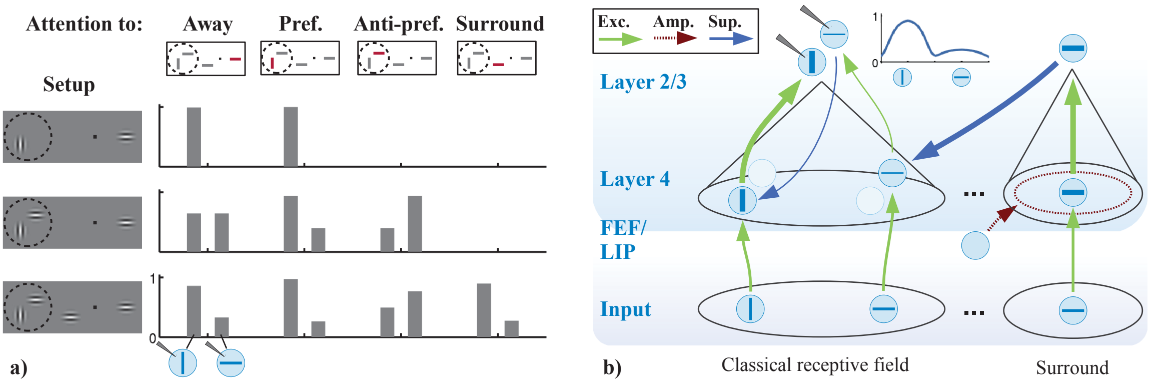


Figure 3.10.: Prediction of biased competition with surround suppression. **a)** The setup consists of a preferred (vertical grating) and anti-preferred stimulus (horizontal grating) inside the classical receptive field (cRF), plus an anti-preferred stimulus in the surround. Three setup variations were simulated: the preferred stimulus alone (first row), biased competition alone containing the preferred and anti-preferred stimulus in the cRF (second row), and biased competition with the additional anti-preferred stimulus in the surround (third row). Model responses are illustrated in the condition of attending away or of attending one of the three stimuli. The attended stimulus is marked in red. Responses are not displayed if the attended stimulus is absent in a setup. **b)** Model mechanisms, illustrated in the condition of spatially attending the surround stimulus (last condition in a). The notation of the figure is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness.

3.3.5. Predictions

The model predicts interactions between many of the previously outlined paradigms as the underlying mechanisms will normally interact with each other. We will exemplarily investigate the interactions of biased competition and surround suppression.

The competition between a stimulus pair can be favored for one of them by attending it (biased competition). However, a surround stimulus can also favor one of them by selectively suppressing the other one. An exemplary setup is depicted in Fig. 3.10a, third row. It consists of a preferred (vertical grating) plus an anti-preferred stimulus (horizontal grating) inside the classical receptive field (cRF), and an additional anti-preferred stimulus in the surround (horizontal grating). Stimulus contrast is high and equal for all stimuli. We propose to record from a neuron in area V4 as biased competition and surround suppression effects were reported for this area (Reynolds et al., 1999; Sundberg et al., 2009). As surround suppression strength depends on the stimulus size (Cavanaugh et al., 2002a) and as the suppression is weak in V4 for small stimuli (Sundberg et al., 2009), we use a large surround stimulus to obtain a strong effect (3x of the cRF size). Compared to this, the stimuli inside the cRF are much smaller (1/5 of the cRF size). The interaction between surround and stimuli inside the cRF can be very complex (Sec. 3.3.4). We assume here the most typical case in which only similar stimuli suppress each other.

The model predicts that an anti-preferred surround stimulus decreases the response to an anti-preferred stimulus in the cRF, leading to an increased response of a preferred stimulus in the cRF. We simulated three setups (Fig. 3.10a): the preferred stimulus alone (first row), biased competition (second row), and biased competition with a surround stimulus (third row). For all setups, we simulated the attentional conditions of 1) attend away, 2) attend preferred, 3) attend anti-preferred, and 4) attend surround stimulus. The model uses in all simulations its standard parameters (Tab. 3.1 in Sec. 3.2). The first setup contains the preferred stimulus only, and so illustrates the response without suppression from other stimuli. Unattended, the response is already strong due to the high stimulus contrast, so attending the stimulus increases the response only marginally. The second setup shows the classical biased competition effect, namely the average response to the pair of stimuli if attention is directed away. Furthermore, if one of the two stimuli is attended, its response is amplified and the other one suppressed. The third setup combines biased competition with surround suppression. In the attended away condition, the response of neurons encoding the preferred stimulus is halfway between the response to the stimulus alone and to the pair in the biased competition setup. Complementary, the response of neurons encoding the anti-preferred stimulus is lower. Both responses indicate that the surround stimulus suppresses the anti-preferred stimulus in the cRF. Attending the surround stimulus increases this effect. Attending one of the two stimuli inside the cRF results in the previously shown biased competition effects with a minor influence of surround suppression, illustrating that the amplification has a stronger influence than the surround suppression.

Attending the surround stimulus has the most prominent effect (Fig. 3.10b). In the model, it amplifies the response of neurons encoding this stimulus, which strongly suppress neurons encoding the anti-preferred stimulus inside the cRF. Contrary, neurons encoding the preferred stimulus are not suppressed as surround suppression occurs only between neurons encoding similar features, here the features of the anti-preferred stimuli. In the cRF, the reduced response of those neurons diminishes the feature-based suppression to the neurons encoding the preferred stimulus. The response of the neurons selective for the preferred stimulus is so almost similar as in the condition of presenting the preferred stimulus alone. If the surround stimulus is not attended (attend away condition), the mechanisms operate identical but are weaker. If a stimulus inside the cRF is attended, the mechanisms of biased competition have more influence than surround suppression. Attending the preferred stimulus removes the influence of the weak suppression from the distractor in the cRF, so the response increases slightly. Attending the anti-preferred stimulus amplifies the response of neurons encoding it, which diminishes the influence of suppression. This suppression results from the neurons in the surround as well as from the neurons encoding the preferred stimulus (feature-based suppression). Thus, the accumulated suppression has still a notable influence and the response of those neurons is less increased as in biased competition. This lower response results in less feature-based suppression back to the neurons encoding the preferred stimulus, so their rate is less decreased as in biased competition.

3.4. Discussion

Experimental investigations that focused on the attentional modulation of the neuronal response in visual areas have led to a large set of data. In order to understand the underlying mechanisms of attention, computational models can help if they allow to explain the observations by only a few core mechanisms. In the proposed model, these are amplification, normalization, spatial pooling, and suppression. In the following, we will review the abilities of existing models and the mechanisms they incorporate. Besides, we will discuss limitations of the proposed model. Furthermore, we will discuss the physiological foundation of the approach, especially concerning the implementation of the suppression.

3.4.1. Relation to existing single-area models

The motivation for a unified model of visual attention arises mainly from the facts that each of the more recent existing single-area models can account for only a subset of the data (Tab. 3.2) and that most models do not explain the involvement of underlying mechanism within the cortical microcircuit.

The model of Reynolds and Heeger (2009) is probably the most popular one as it can replicate data from several experiments: modulation of the contrast response function, modulation of the neuronal tuning curve, and biased competition with feature-based at-

Fig. no.	Experiment	Reynolds & Heeger (2009)	Boynton (2009)	Lee & Maunsell (2009)	Ni et al., (2012)	Spratling (2008)	Wagatsuma et al. (2013)
3.2, 3.3a	Biased competition with spatial att.			+	+	+	
3.3a,b	Feature pooling			+	+		
3.3c-e	Selectivity vs. sensory interaction						
3.3f	Contrast dependency of suppression						
3.4	Biased comp. with feature-based att.	+				+	
3.5	Contrast gain	+	+				
3.6	Response gain	+	+				
3.7	Scaling	+	+			+	+
3.8a	Sharpening	+	+			+	+
3.8b	Tuning curve in biased competition	imprecise fit					
3.9a,b	Surround suppression						

Table 3.2.: Capabilities of existing single-area models to account for a particular attention experiment. The symbol ’+’ denotes cases were the publication shows simulation results for a certain experiment.

tention. Essential ingredients for its abilities are divisive normalization (Carandini and Heeger, 2012) accounting for all contrast related properties, and a generic attention plus a suppression field to simulate amplification and suppression respectively. Reynolds and Heeger (2009) state that they made no assumption how and over which connectivity these interactions are carried out in the cortical microcircuit. This improves the flexibility of the model, however it is a drawback in terms of specific predictions about the putative implementation within the microcircuit. As compared to our model, their model has no clear definition of a receptive field as it lacks the spatial pooling operation in layer 2/3. So it cannot replicate experiments where spatial attention is directed to only a part of the receptive field, e.g. biased competition with spatial attention and selectivity versus sensory interaction (Tab. 3.2, rows 1 and 3).

To recapitulate, the biased competition experiments with spatial attention (Reynolds et al., 1999) use two stimuli, which are placed within a receptive field at different positions. In our model, a spatial pooling operation is realized by excitatory connections from layer 4 to layer 2/3. By these converging connections, the neuronal responses of a neuron in layer

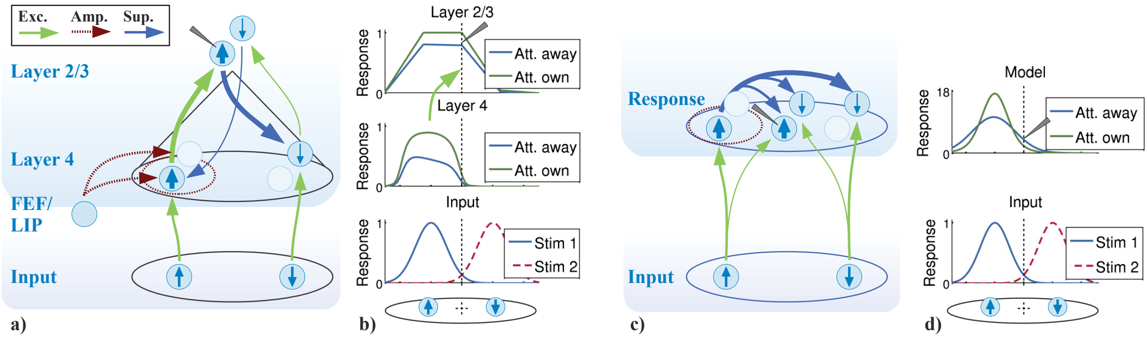


Figure 3.11.: Comparison of our model with spatial pooling (a, b) with the model of Reynolds and Heeger (2009) lacking pooling (c, d), in a spatial biased competition experiment. **a)** The proposed mechanistic cortical microcircuit of attention. Notation is similar to Fig. 3.2. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness. **b)** For our model with spatial pooling, the responses of the input layer, layer 4, and layer 2/3 are illustrated over all location within the receptive field. The responses of the upward moving stimulus are denoted by blue (unattended) and green colors (attended own), and the downward one is shown in red, but for simplicity only in the input. The model assumes symmetrical receptive fields, thus the recorded cell is located in the center, denoted by a dotted line. **c)** Model without spatial pooling proposed by Reynolds and Heeger (2009). Notation is identical to a). To facilitate a comparison with a), the mathematical formulation of the model is illustrated as a neuronal circuit. **d)** For the model without spatial pooling, responses of the input and the recorded layer. Notation is identical to b).

4 stimulate the recorded neuron (Fig. 3.11b, green arrow) which defines the receptive field by its bottom up connectivity. Thus, any amplification caused by spatial attention to a part of the receptive field will amplify also the recorded neuron (blue versus green curve in layer 2/3). Using this mechanism, the proposed model replicates the physiological data (Sec. 3.3.1), i.e. that the attended response is higher as the unattended one.

Contrary to our model, Reynolds and Heeger (2009) (Fig. 3.11c, d) use a spatial broadening of the stimulus instead of a pooling operation. Indeed, in the unattended condition, this approach is effective as the neuron at the center location is activated (Fig. 3.11d, blue). Attending the stimulus location amplifies also the neuronal response at the location of the stimulus (Fig. 3.11d, green). However, this amplified response induces a stronger suppression at the other locations, thus the response will decrease at the recorded neuron located at the center. This decrease is obviously in contrast to the observed amplification in biased competition (Lee and Maunsell, 2010a), as shown in Fig. 3.2 (Sec. 3.3.1). However, if the model of Reynolds and Heeger (2009) would be extended by a spatial pooling layer, it could theoretically simulate this biased competition paradigm. In summary, any attention experiment that requires to spatially attend only to a part of the receptive field and to record at another location, cannot be simulated by a single layer neuronal model.

Others (Boynton, 2009; Lee and Maunsell, 2009, 2010a,b) published simpler computational models based on divisive normalization. The model of Boynton (2009) is primary intended to simulate the modulation of the contrast response function. Thus, their model has a contrast and feature dimension, but no spatial dimension and no attention or suppression field. Attention is modeled in a very abstract parametric fashion and not by an intrinsic mechanism, rather it is applied via three independent parameters: the multiplicative scaling of a tuning curve, a baseline gain, and a shift of the tuning curve to the left. These parameters allow to fit the model to data with respect to the attentional modulation of the neuronal tuning curve and contrast response function (Tab. 3.2), but make it unlikely that the model can account for other data. In particular, it lacks any suppression mechanism.

Another simple model based on divisive normalization is from Lee and Maunsell (2009) (also used and described in Lee and Maunsell (2010a,b)). Their model simulates a response based on a driving input and a normalization, which may involve one or two stimuli. The normalization term in the denominator of the neuronal response function operates as suppression. It is simulated via an exponential function based on the contrast and a nonlinearity, a somewhat different type of normalization compared to Carandini and Heeger (2012). A stimulus is only encoded by a single rate variable instead of a neuronal population, so the model lacks tuning curve characteristics. To overcome this limitation, they model only the interactions between two stimuli. More precisely, they firstly calculated the tuning curve for a single stimulus from the physiological data without using their model. Secondly, they simulate the interactions by passing the previously calculated tuning curves of the two stimuli to their model, which then modulates and normalizes the responses (Lee and Maunsell, 2010a). Due to this procedure, their model can explain the change in the neuronal response when a second stimulus is added in the receptive field, but not the response towards the single stimulus itself. In summary, they simulate biased competition and feature pooling, both with spatial attention (Tab. 3.2, rows 1 and 2). The research group of Maunsell uses in Ni et al. (2012) a new model to explain their findings that neurons are modulated differently strong by attention. Their approach is called tuned normalization, and it is more in line with the approach of divisive normalization. Otherwise, this model incorporates the same ideas as the previous one (Lee and Maunsell, 2009, 2010a,b). In summary, their model contains all necessary mechanisms (amplification, normalization, and suppression) except spatial pooling to simulate all considered attention experiments (Tab. 3.2), so it could theoretically simulate more experiments as the published biased competition experiments.

Spratling (2008) proposed an attention model that has a higher biological plausibility than the previously reviewed models because it explicitly includes neural connections. The model contains two stages, possible representing two different brain areas (Spratling, 2012), whereby each stage contains two sublayers: a prediction and an error layer. The prediction layer sends an expectation signal to the error layer, so the error layer can minimize the reconstruction error. Attention signals are deployed to the prediction layer, which are then transported to the error layer. Spatial attention is applied to stage 1 and feature-based attention to stage 2. Thus their and our model have a similar structure because both use two layers to simulate a single cortical area, but their model is fundamentally different in all other aspects. Spratling (2008) can replicate a few attention paradigms with his model: modulation of the contrast response function, modulation of the neuronal tuning curve, and the decrease of reaction times (Tab. 3.2). A different version of his model (Spratling, 2010) can replicate additionally the data of MacEvoy et al. (2009) (competition and feature pooling without attention) and of Cavanaugh et al. (2002a) (surround suppression without attention). However, this model variant does not contain attention signals, and so we do not include it into our list of attention models (Tab. 3.2).

An attention model that has the advantage to precisely simulate the cortical microcircuit is Wagatsuma et al. (2013). Their model is a revised version of an existing model of the V1 microcircuit (Potjans and Diesmann, 2012) consisting of layers 2/3, 4, 5, and 6. Their model uses spiking neurons to allow a precise modeling of the temporal effects of attention. Thus the model is focused on replicating such effects like oscillations, but it can also replicate two experiments considered here (Tab. 3.2): the scaling and the sharpening of neuronal tuning curves. However, due to the high complexity of the neuronal circuit, it cannot easily explain which mechanisms are necessary for a particular attention effects. Moreover, the range of simulated experiments is presently too small to draw general conclusions about the role of connections in the simulated microcircuit. There exist also a few other models based on spiking neurons (Ardid et al., 2007; Buia and Tiesinga, 2008; Hugues and José, 2010), but with much less realistic modeling of the microcircuit and no further advantages.

3.4.2. Model limitations

The proposed model of visual attention focuses on the explanation of single unit recordings by means of a mechanistic cortical microcircuit of attention. Thus, the model covers the information processing necessary for replicating the neural firing rates emerging in these experiments, but it simplifies all other aspects as much as possible to keep the model simple and focused. For example, the model contains only layer 4 and layer 2/3 of the cortical microcircuit (Douglas and Martin, 2004) and only the necessary part of its connectivity.

Furthermore, to allow an easy visualization of neuronal responses, we model only a single feature dimension, and a single spatial dimension instead of two spatial retinotopic dimensions. We also do not simulate any feature integration or learning in the visual cortex, hence all layers use the same features.

As the model focuses on the simulation of a cortical microcircuit of attention, experiments involving multiple areas cannot be simulated. In previous studies, we focused on the interaction of the visual cortex with the frontal eye field (Hamker, 2005a) to address the temporal dynamics of attention and the interplay between feature-based and spatial attention. Moreover, splits of spatial attention have been addressed (Zirnsak et al., 2011a). Furthermore, attention effects operating simultaneously on multiple levels cannot be simulated as this case might also involve multiple areas. One exemplary experiment is object-based attention in the form where a subpart of an object is attended which induces attention to the whole object and to other subparts (Mitchell et al., 2004; Reynolds and Heeger, 2009).

Furthermore, temporal effects of attention are beyond the grasp of this model. For example, attention modifies the pattern of neural oscillations (Jensen et al., 2007; Grossberg and Versace, 2008), decreases the transmission delays (Sundberg et al., 2012), decreases the variability of the neuronal response (Mitchell et al., 2007) or enhances the temporal coupling of neurons (Gregoriou et al., 2009). These findings may be crucial to understand the full neural implementation of the proposed cortical microcircuit of attention and would require a different level of implementation detail.

3.4.3. Physiological origin of suppression

At present, it is not accurately known (Reynolds and Heeger, 2009) how the basic mechanisms of attentional processing, i.e. amplification, normalization, and suppression, are implemented in the cortical microcircuit of a visual area (Douglas and Martin, 2004). As the mechanics of amplification and normalization in our model are inspired by Reynolds and Heeger (2009) and Carandini and Heeger (2012), we would like to refer to this literature about the physiological foundations of the first two mechanisms. Very recent research shows also additional possibilities, for example Brosch and Neumann (2014) have proposed an implementation of amplification via interneurons. Concerning the suppression mechanism, our model proposes a new and more concrete implementation, so we will substantiate this point by relating our model to the suppression literature.

In our model, suppression is implemented for simplicity only via connections from layer 2/3 to layer 4. In the cortical microcircuit (Potjans and Diesmann, 2012), such connectivity could be realized by a chain of anatomical connections: excitatory connections from layer 2/3 to layer 5/6, excitatory connections from layer 5/6 to layer 4, and inhibitory connections within layer 4. However, we are aware that there exist also other possibilities to implement the suppression in the cortical microcircuit (Potjans and Diesmann, 2012). We have chosen layer 2/3 as starting point for the suppression as layer 2/3 is the entry point for feedback connections from higher areas (Douglas and Martin, 2004). Feedback seems to transport suppressive signals from higher to lower areas.

In principle, suppressive signals can be transmitted via feedforward, lateral, or feedback connectivity (Gilbert, 1998). Feedback connections seem to be involved in surround experiments. The reviews of Carandini and Heeger (2012) and Gilbert (1998) rule out feedforward connections, Bair et al. (2003) concludes that suppression of neuronal responses originates from feedback signals, and Angelucci et al. (2002) found a correspondence between the spatial extent of feedback connections and of surround suppression. Lateral connections seem mostly to be involved in low contrast experiments. Angelucci et al. (2002) found that the receptive field size of low contrast stimuli corresponded to the extent of lateral connections. Hunt et al. (2011) found that excitatory lateral connections typically resemble large structures available in natural scenes, like edges or circles, denoted as co-linearity or co-circularity. A study by Cass and Spehar (2005) links these two facts together: they found that the range of collinear contrast facilitation is consistent with long-range horizontal transmission. In conclusion, suppression mediated by feedback from higher areas can be easily incorporated into our model.

3.4.4. Relevance of the model for psychophysical experiments

The proposed attention model of the cortical microcircuit can explain the data of several physiological attention experiments, but the model mechanisms are also very relevant for psychophysical experiments.

One psychophysical aspect of attention is its selective processing, because it selects task relevant stimuli under many distractors (Carrasco, 2011). The selection might be necessary to not overload the visual systems limited capacity (Todd et al., 2005), or to save energy as the brain's metabolic cost depends mostly on the neuronal activity (Carrasco, 2011). In the model, such a selection can be achieved by amplification of the target and by suppression of distractors. Visual search (Lee and McPeek, 2013) is a typical experiment involving such a selection. Suppression is also observed in crowding (Whitney and Levi, 2011) where the visibility of a stimulus is reduced if it is shown with nearby stimuli. One influential factor for the reduced visibility may be our proposed suppression mechanism(s).

Besides the selection of stimuli, attention increases the behavioral acuity of subjects by increasing contrast sensitivity, by decreasing reactions times, or by reducing external noise (Carrasco, 2011). The increased contrast sensitivity is also visible in the simulated physiological data as attention causes an appropriate modulation of the contrast response function (Sec. 3.3.2). Noise reduction can be realized by amplifying the neuronal response of the target, and by suppressing all other neuronal responses. Therefore, many psychophysical aspects of attention will rely on the mechanisms proposed in our model.

3.4.5. Influence of model parameters

Our model contains several free parameters that have been tuned to fit different data sets. However, the model should not be dominated by the particular setting of its parameters. Hence, the model should be able to qualitatively fit the data with a fixed parameter set, and parameter variations should only be necessary for a quantitative fit. The model includes parameters (Tab. 3.1) to modify the contrast response function (parameters v^{E}, p^{E}), to scale the influence of amplification ($v^{\mathrm{FEF\text{-}L4}}$, $v^{\mathrm{PFC\text{-}L2}}$) or suppression ($v^{\mathrm{FEAT}}$, p^{FEAT}, v^{SUR}, p^{SUR}), to add an un-modulated baseline (β), and to change the neuronal tuning in the feature space (tuning curve). We evaluated if the model can qualitatively fit the data by fixing all parameters to a reasonable value, denoted as standard parameter set (Tab. 3.1, row 'std'). Only the neuronal tuning curve was allowed to be freely chosen because of the heterogeneous response characteristics of neurons in different cortical areas.

The results, illustrated in the Appendix A, show a satisfying qualitative fit for nine of the twelve experiments. In two experiments (Sundberg et al. (2009), Fig. 3.9b; Williford and Maunsell (2006), Fig. 3.6), our results reproduce all significant effects of the original publication, but differ notable from the data. In one experiment (Cavanaugh et al. (2002a), Fig. 3.9a), a significant effect cannot be reproduced. The deviations result in all three data sets from an inappropriate strength of the surround suppression.

We found that the surround suppression strength (v^{SUR}) and non-linearity (p^{SUR}) vary strongly among data sets (Tab. 3.1). In the V4 data of Sundberg et al. (2009) (Fig. 3.9b), the surround suppression is weak which we quantitatively fit by setting $v^{\text{SUR}} = 0.55$. Contrary, the V1 data of Cavanaugh et al. (2002a) (Fig. 3.9a) was best fitted with a strong surround suppression of $v^{\text{SUR}} = 0.8$. Besides these two data sets, the surround suppression mechanism is also used for suppression from close, adjacent neurons in four data sets (Fig. 3.3b, 3.6, 3.7, 3.8a). These were fitted optimally by $v^{\text{SUR}} = 0.5, 1.0, 0.75, 0.5$. Thus, the standard value for v^{SUR} is chosen as a compromise of all these values: $v^{\text{SUR}} = 0.5$.

The strength of the non-linearity p^{SUR} varies as well. In the data sets of Sundberg et al. (2009) and Cavanaugh et al. (2002a), $p^{\text{SUR}} = 2.5$ and $p^{\text{SUR}} = 2.0$ results in the best fit. However, in experiments with suppression from close, adjacent neurons, disabling the non-linearity by setting $p^{\text{SUR}} = 1$ results in the best fit. The diversity might depend on factors such as if the suppression originated from neurons in the far (Fig. 3.9a, 3.9b) or close surround (Fig. 3.3b, 3.6, 3.7, 3.8a). We have chosen as standard value $p^{\text{SUR}} = 1$ as this value allows to fit more data sets. Unfortunately, this standard value prevents that the model can reproduce the significant increase of receptive field at lower contrasts in the data of Cavanaugh et al. (2002a), as discussed in the supplementary materials. Yet, the model still shows suppression from the surround at all contrasts. As a side effect, high values for p^{SUR} decrease the influence of the suppression as well.

Therefore, the fits of Sundberg et al. (2009) and of Williford and Maunsell (2006) differ notable as the suppression has an incorrect strength. The standard values of $v^{\text{SUR}} = 0.5, p^{\text{SUR}} = 1.0$ results in a suppression which is too strong for the experiment of Sundberg et al. (2009), which is optimally fitted by $v^{\text{SUR}} = 0.55, p^{\text{SUR}} = 2.5$, and is too weak for the experiment of Williford and Maunsell (2006), which is optimally fitted by $v^{\text{SUR}} = 1.0$, $p^{\text{SUR}} = 1.0$.

	1) Amplification	2) Normalization	3) Spatial pooling	4a) Feature-based suppression	4b) Surround suppression
Biased competition with spatial att.	+		+	+	
Biased comp. with feature-based att.	+			+	
Contrast gain	+	+			
Response gain	+	+			+
Scaling	+				
Sharpening	+			+	
Surround suppression	+				+

Table 3.3.: Overview of the primarily involved mechanisms in each experiment. The related explanation of a mechanism in a particular experiment can be found in Sec. 3.3.

In summary, the standard model can replicate the main effects of all data sets with minor deviations regarding surround suppression effects. Thus, all parameters could be fixed for a qualitative fit except the surround suppression parameters $v^{\mathrm{SUR}}, p^{\mathrm{SUR}}$. Therefore, we conclude that primarily the model mechanisms, and not the free parameters are responsible to account for the data sets.

3.5. Conclusion

In this work, we proposed a new mechanistic model of the cortical microcircuit to explain the neuronal response modulation caused by visual attention. For this purpose, we unify existing proposals of attention into a new neuro-computational model, and underline psychophysical and physiological concepts of attention with constraints from neuroanatomy. We found that many visual attention effects can be explained by a few mechanisms: 1) amplification of the response of a single neuron, 2) divisive normalization of this response, 3) spatial pooling within the receptive field, and 4) suppression between neurons encoding different features or different locations. By these mechanisms (Tab. 3.3), we are able to explain the following broad variance of physiological single-unit observations by a single model: biased competition with spatial or feature-based attention; modulation of the contrast response function resulting in contrast or response gain; modulation of the neuronal tuning curve resulting in scaling or sharpening; and modulation of surround suppression.

Moreover, our model can predict attention effects and mechanisms in novel paradigms, which we demonstrated by combining the paradigms of biased competition and surround suppression.

The current model could be used as a core element in larger models of visual attention that include multiple areas in the visual cortex and control structures like FEF/LIP and PFC/7a. Such a system-level model could simulate very well psychophysical experiments as already demonstrated by a previous system-level model of attention of our group (Hamker, 2005a,b) for receptive field dynamics (Hamker and Zirnsak, 2006), spatial compression (Zirnsak et al., 2010), or the split of spatial attention (Zirnsak et al., 2011a). Besides psychophysics, our model may also inspire computational vision models in robots, e.g. Antonelli et al. (2014), or in other computer vision applications like object localization. In the next two chapters of this doctoral thesis, we will precisely realize such ideas.

4. Object localization with a neuro-computational model of visual attention

Abstract A very relevant task in computer vision is to localize a target object in a scene. Humans perform this task with the help of visual attention, but the process is not fully known and neuro-computational models have been rarely used in realistic problems. Computer vision systems often use attention as a spatial pre-selection stage (saliency models), but in the brain, the processing seems to spawn rather a top-down control network, modulating neuronal activity for the current task. Therefore, we like to propose a more general view of attention as a cognitive, holistic control process, and like to illustrate, with a newly-developed model, again the attentional processing in object localization to further spread this knowledge to the community. According to our work, the brain localizes objects as follows: Attention is deployed to a target object, precisely to its type (feature-based attention). This results in top-down signals from the prefrontal cortex to visual cortices, which enhances neurons encoding views of the target object and suppresses others. Spatial cortices rely then on this biased activity pattern, and select by means of a recurrent processing the target location. The novel model was developed here to enhance previous neuro-computational models with sophisticated concepts: the neurophysiological-grounded mechanisms (Chap. 3), a new frontal eye field, and learning of object representations. At first, we demonstrate with the novel model that neuro-computational models are applicable also to realistic computer vision problems, because currently, only a few of such models have been applied to real-world object localization tasks and merely with a low number of objects (2-16). For this, we ran our model on a large setup with 100 objects, 1000 scenes, and three background classes (black, white-noise, and real-world). The high number of objects is possible due to the learned object descriptors. The model achieves a localization accuracy of 92% at black backgrounds, at which it was learned. A generalization to white-noise and real-world backgrounds changes the accuracy to 71% and 42% respectively. Secondly, we shed more light on the attentional processing in the task. From neurophysiology, we know that attention can amplify (amplification mechanism) or suppress neuronal responses (suppression). However, the precise roles of these mechanisms in object localization have remained vague. We focus here on feature-based attention, and

found: The feature-based amplification mechanism performs the target neuron enhancement, and thus represents the target in the visual cortex for the top-down control. The feature-based suppression removes neuronal noise originating from other objects or the background, but is irrelevant for the behavioral performance in this task. To conclude, we shed more light on the role of attention in object localization and demonstrate that neuro-computational attention models are applicable to realistic computer vision tasks.

This chapter has been published in parts as: Beuth and Hamker (2015b).

4.1. Introduction

A typical task for computer vision systems is to localize an object of interest in a scene. For example, a robot may contain such an object localization system to localize the relevant objects for its current task, so it can interact with them (Antonelli et al., 2014). Another usage may be for video annotation, where often the question occurs where are certain persons or other object of interests in the video image (Ritter, 2013). The task is daily executed as well by us humans, for example when we have to localize a pen at our office desk for writing. Thus, it is from interest in the field of psychology too, where the task is known as guided visual search (Wolfe, 1994). We define here an object localization task as any task where the target object is given via its type, and its location is unknown.

Humans perform this task with the help of visual attention, as shown by the psychological research about guided visual search (Chap. 2). This led over the years to different approaches utilizing visual attention (Chap. 2.4): bottom-up saliency models, top-down saliency models, and cognitive holistic models. The first two groups, the bottom-up and top-down saliency models (for good reviews see Borji and Itti (2013), Filipe and Alexandre (2013), Frintrop et al. (2010)), are the classical computer vision approach using attention: The idea is to utilize attention as a spatial pre-selection mechanism for a subsequent object-classification stage. The concept was inspired from the classical psychological view of attention as a selection process (Carrasco, 2011). In the bottom-up approach, attention emerges in a bottom-up manner from the scene and represents locations with large amount of information, called conspicuity. The approach filters the image with simple feature descriptors, like edges or color contrasts, to obtain several feature maps. Afterwards, these are normalized and integrated together to form a saliency map, representing the conspicuity of the image. A region of interest (ROI) is selected around the most salient location, and a sophisticated classifier tries to recognize the target object in the ROI. If the target is not found, the next ROI is processed. A typical example of the approach is the Neuromorphic Vision Toolkit (NVT) from the group of Itti (Itti et al., 1998; Itti and Koch, 2001). As advantage, the approach is very fast as the computational-expensive classifier is only executed a few times. It is also capable to work with real-world scenes as many studies showed applications in this domain (Chap. 2.4.1).

The drawback of this bottom-up approach is that non-salient objects are not detected (Frintrop et al., 2010), leading to the development of top-down approaches. The most authors extend the bottom-up approach by adding top-down attention to the feature maps in the saliency model. This top-down signal amplifies features belonging to target objects, resulting in an enhanced saliency for these objects. Afterwards, ROIs are again selected from this "biased" saliency map, and objects in these regions are recognized. A typical example of top-down attention systems is VOCUS from Frintrop (Mitri et al., 2005; Frintrop et al., 2004; Frintrop, 2006), used by Mitri et al. (2005) in a robot soccer scenario. Many other models use the concept too as outlined in Chap. 2.4.1, for example Borji et al. (2009), and Navalpakkam and Itti (2006).

However, attention in the primate brain seems to work differently than in the saliency models, whereby its precise role is still under debate. Recent neuroscience studies showed that the attentional processing spawns a top-down control network modulating neuronal activity (Miller and Buschman, 2013). The network originates from the prefrontal cortex (PFC) which represents task sets (Sakai, 2008), and targets the whole visual cortex in parallel (Miller and Buschman, 2013). Hence, attentional processing modulates neuronal activity in the whole visual cortex for the current task, thus attention is not solely a spatial preselection stage as in the saliency models (see Chap. 2.4 for details). Due to these recent neuroscience findings, we like to propose (Sec. 4.2.3) a novel, more general view of attention than the classical view of a selection process (Carrasco, 2011): a cognitive, holistic control process, which tunes the visual system for the task at hand. Holistic means in our terms that the control process runs in parallel within the whole visual cortex, and modulates everywhere neuronal activity for the current task. As the saliency models are much more popular than the neuro-computational models in the computer vision community (as shown in Chap. 2.4), many may not know the differences between the attentional processing in the saliency models and the human brain. Therefore, we like to illustrate again the processing of visual attention in object localization for the computer vision community to further spread this knowledge.

Therefore, we denote all models that simulate the attentional processing of the primate brain as holistic. Such models have been already applied to object localization, namely in the phenomenon of guided visual search. Yet, the majority have been used in tasks with simple and artificial stimuli (Bruce and Tsotsos, 2011; Deco et al., 2002; Deco and Rolls, 2004; Lanyon and Denham, 2004; Rodríguez-Sánchez et al., 2007; Tsotsos et al., 1995; Wolfe, 1994), and thus are probably unable to operate with whole objects or real-world scenes. Only a few models have been applied to localization tasks with whole objects and real-world scenes: Antonelli et al. (2014), Beuth et al. (2010), Chikkerur et al. (2010), Hamker (2005b), Walther and Koch (2007). Other models can work under these conditions too, but focus on modeling the visual cortex instead of attentional processing (ART, Carpenter et al. (2005), Grossberg (2013); VisNet, Rolls (2012)), thus we do not consider them here. Holistic models have two advantages over the top-down saliency models (Chap. 2.4.3): Firstly, selection and recognition operate in separate stages in the saliency models, but should be intertwined as they depend on each other (Antonelli et al., 2014; Frintrop et al., 2010). In the holistic models, attention controls both processes in parallel. Secondly, saliency models select a region only via simple features, whereas the holistic models invoke sophisticated object representations in high-level visual areas.

In the following, we explain the holistic models and their attentional processing exemplary on the model of Hamker (2005b). We will build our model on it as it has, from the five real-world holistic models, the best physiological and neuroanatomical grounding (for its grounding, see Hamker (2005a)). The model simulates attentional processing via a complex interaction of brain areas, mainly between the prefrontal cortex (PFC), higher visual areas (IT, V4), and the frontal eye field (FEF). The FEF is an area involved in spatial processing. Hamker (2005b) explains the attentional processing in object localization as follows. Our explanation will largely follow this one as our model will be based on this model. When a target object should be localized, feature-based attention is deployed to the target object. This activates cells encoding the target object's type in PFC, which then biases the high-level visual areas for the target by amplifying there target-feature encoding neurons, while others neurons are inhibited by local connections. Afterwards, the FEF extracts from this activity pattern a map with potential target locations, and projects back to the visual areas, forming a reentrant processing loop. The loop focuses neuronal activity to the best location from the potential ones and in such a way selects the location of the target. The process to focus activity to a single location is seen as the emergence of spatial attention. The spatial attention loop synchronizes continuously the information about the target location between the FEF and the visual areas, controlling both selection and recognition, and in such a way allowing the parallel execution of both processes. Therefore, attention initially controls the high-level visual cortices for a task-relevant target via feature-based attention, and then spatial attention emerges between these visual cortices and the FEF, which controls the system regarding spatial aspects and selects the target location as final outcome of the localization process.

The model of Hamker (2005b) has some weaknesses, which we would like to improve. Firstly, the neuronal mechanisms in Hamker (2005b) are able to account for some neurophysiological effects of attention (Hamker, 2004b, 2005a; Hamker and Zirnsak, 2006; Hamker et al., 2008), but not for so many effects as the microcircuit model (twelve different effects, Chap. 3). Hence, we incorporate the microcircuit model into our novel system-level model. Secondly, we improve the model's frontal eye field (FEF). The old FEF did not consider the anatomical finding that the FEF visuomovement cells project back to V4, and not the FEF movement cells as implemented (Pouget et al., 2009). This connectivity was revised in an earlier work of the author (Zirnsak et al., 2011a), hence the FEF is updated based on this work. Finally, we introduce a learning procedure to create high-level object representations. The model of Hamker (2005b) does not contain learning, and encodes an object as a simple set of low-level features. In contrast, the primate brain employs much more high-level object representations, like cells encoding a view of an object (Logothetis et al., 1995). Such cells have been learned by trace learning in previous models of the author (Antonelli et al., 2014; Beuth et al., 2010), thus we employ this method also here. This advancement makes it possible to encode a large number of objects, which is a necessity for our benchmark.

The existing holistic real-world models have several weaknesses which we would like to resolve with our new model. Firstly, they have been evaluated only in setups with a few object categories (2-16), raising doubts about the robustness of the approach. To nullify this issue, we use a large and realistic setup with 100 objects, 1000 scenes, and three different backgrounds classes (black, white-noise, and real-world).

Secondly, the existing studies have not clarified the role of the known neuronal mechanisms of attention (Chap. 3) in object localization. Thus, we like to investigate this point to better understand the attentional processing in object localization. From neurophysiology, we know that attention can amplify (amplification mechanism) or suppress neuronal responses (suppression). The existing studies have also presented already explanations for the attentional processing in object localization (for example, see Hamker's explanation), but typically use several neuronal mechanisms to simulate this complex processing, hence the role of an individual mechanism within this processing remained unclear. Due to this, it is also still unclear how much a particular mechanism contributes to the object localization performance, and how the mechanisms realize the holistic control process. As neuronal attention mechanisms characterize the implementation level of attention in the brain, such knowledge would be very preferable for a future implementation of attention into computer vision systems. Therefore, we like to analyze the individual role of each mechanism. For this purpose, we first explain each mechanism's role in object localization (Sec. 4.3), and afterwards conduct an individual evaluation (Sec. 4.4) by disabling each mechanism separately and then comparing neuronal activity and behavioral performance of the impaired system to the full system. In such a way, we can precisely investigate the role of an individual mechanism in object localization. As all holistic and many top-down saliency models contain mechanisms to amplify or to suppress neurons, we focus on these two mechanisms. We center further on the mechanisms in their feature-based variants as feature-based attention seems strongly relevant for object localization.

Furthermore, our analysis links the neurophysiological effects of attention to behavior. As our mechanisms are based on neurophysiological data sets of attention and hence every mechanism is associated with several neurophysiological effects, we hereby also work out which neurophysiological aspects of attention are relevant for object localization. This will link neurophysiology to behavior. Therefore, our analysis will give a deeper explanation of the attentional processing at neuronal level in object localization, which facilitates the incorporation of attention into future computer vision systems and links the neurophysiological and behavioral aspects of attention.

Finally, surprisingly few studies have quantified the performance boost through attention, thus it is currently unclear to which degree attention really benefits object localization. The property has been assessed in the following models by evaluating them with and without attention: Borji et al. (2009), Chikkerur et al. (2010), Navalpakkam and Itti (2006), Walther and Koch (2007). As we already evaluate our model similarly, we will give a survey of the achieved boosts in our and other models.

After this introduction, the remainder of this chapter is structured as follows: In section 2, we will elucidate our novel model of attention and illustrate its processing. The attention mechanisms will be explained in section 3 and evaluated individually in section 4. Afterwards, section 5 will discuss our work, while section 6 will conclude it with a focus on the role of the attention mechanisms in object localization.

4.2. Model

We will give first an overview of the model before elucidating our proposal of attention as cognitive, holistic control. Afterwards, we will illustrate the novel learning of object descriptors, and outline the mathematical description of the model.

4.2.1. Overview

Our novel system-level model of attention (Fig. 4.1) consists of lower visual area (V1), a higher visual area (HVA), the prefrontal cortex (PFC), and the frontal eye field (FEF).

Monocular real-world images are processed by the first stage V1 (primary visual cortex), encoding simple visual features like local color contrasts and the orientation of edges. The neurons are organized in three channels, each containing eight features of which three are displayed in Fig. 4.1. The channels encode:

L-M: Local color contrasts between the colors red/orange and green. Colors are perceived by three types of cones in the retina, named after their wavelength sensitivity: large wavelength cones (L) react to red/orange, middle (M) to green, and short (S) to blue (Bowmaker and Dartnall, 1980; Moroney et al., 2002). The L-M cells encode a differential signal between the responses of L and M cones. Such neurons were found in the lateral geniculate nucleus (LGN) as reported by Wiesel and Hubel (1966) (type I neurons) or by Gegenfurtner (2003) (parvocellular layers). We assume that this information is still present in V1.

LM-S: Local color contrasts between the colors yellow and blue, represented by a combined large and middle (LM) wavelength signals, and a short (S) wavelength signal. These neurons were found in the LGN (Type II in Wiesel and Hubel (1966); koniocellular layers in Gegenfurtner (2003)), and we assume again that this information is still present in V1.

O: Eight differently orientated edges. Responses are obtained by converting the image to gray-scale and afterwards filtering it with a Gabor filter. Hubel and Wiesel (1962) found that V1 neurons respond very strongly to such oriented edges, and Jones and Palmer (1987) show that they can be modeled with a Gabor filter.

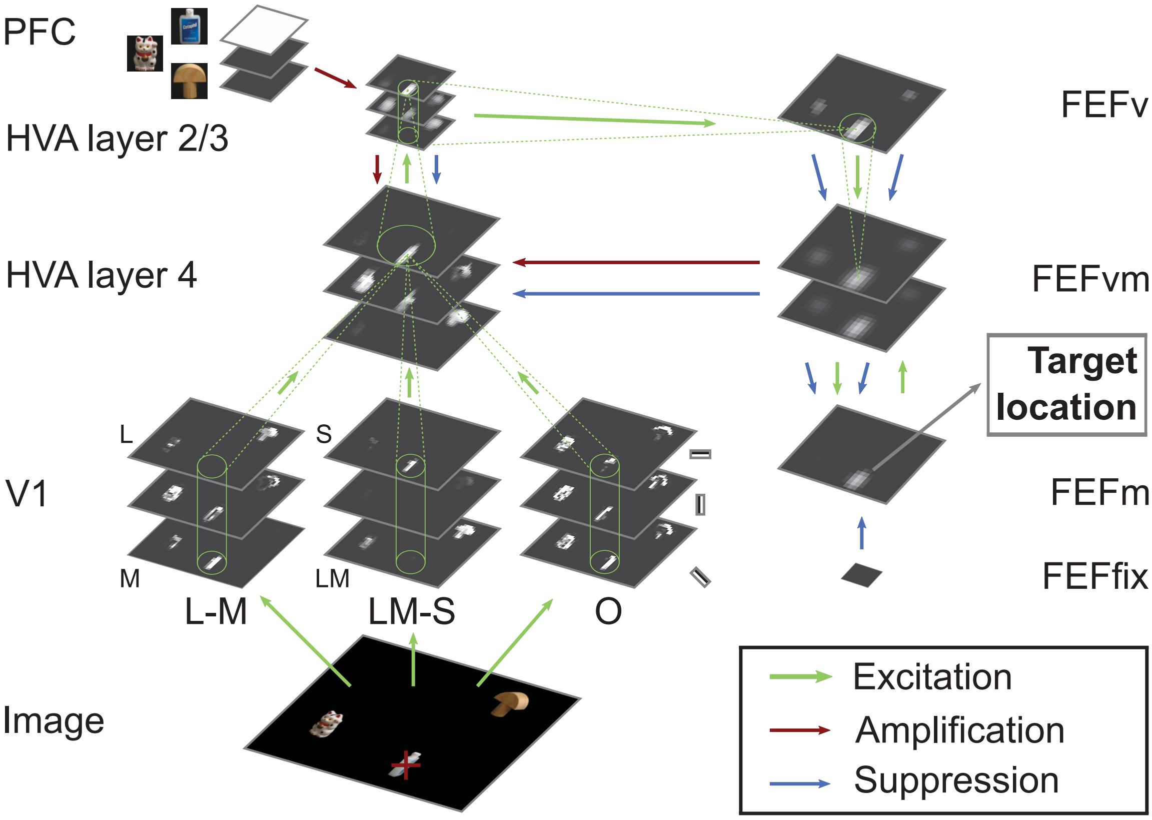


Figure 4.1.: The novel system-level model of visual attention. The processing is illustrated at the task to localize the "bottle", indicated by the red cross. Neuronal responses are denoted by brightness. Firstly, the image is processed by a model of the primary visual cortex (V1), encoding oriented edges (O), red-green (L - M), and yellow-blue color contrasts (LM - S). The "bottle" is mainly characterized by V1 responses in the M feature plane, the S feature plane, and the vertical edge plane. From the V1 activity pattern, a higher visual area (HVA) recognizes object views. For example, the depicted side-view of the bottle is encoded in each top plane of HVA layer 4 and 2/3. The task implies feature-based attention towards the target object, simulated by activating a target encoding neuron in the prefrontal cortex (PFC). This results in a feature-based amplification signal to HVA, increasing the responses of target neurons in both HVA layers. In parallel, feature-based suppression from layer 2/3 to layer 4 decreases neuronal responses towards distracting objects. The frontal eye field (FEF) extracts potential target locations (FEFv) from the so modulated HVA responses. The next layer FEFvm favors neuronal activity at one of these locations. This information is projected back to HVA, spatially amplifying the target location and suppressing other locations. Spatial attention emerges from this recurrent loop between HVA and FEF, leading over time to the selection of a single target location. Finally, an eye movement is planned towards this location (FEFm), indicating the outcome of the localization process.

Within each V1 channel, the neurons are organized feature-wise in planes and each plane has the same spatial arrangement as the image (retinotopic organization). Therefore, a particular V1 neuron will be activated if its preferred feature is located at the retinal location underlying its receptive field.

The next stage HVA models a higher visual area, comparable to high-level visual cortices like V4 or IT. In this object localization setup, HVA contains neurons that represent a single view of an object, similar to the cortical view-tuned cells found in area IT (Logothetis et al., 1995). Neurons are again organized feature-wise in planes and retinotopic. HVA is simulated by our microcircuit model of attention, which is neurophysiologically well founded as it accounts for a large body of neurophysiological attention data sets (Chap. 3). The model simulates a cortical area via two layers, representing the cortical layers 4 and 2/3. In HVA, layer 4 recognizes object views, and layer 2/3 pools spatially over a particularly region of layer 4 cells to increase spatial invariance. The activity of such an object-view neuron in HVA layer 4 is calculated by a weighted sum of its V1 inputs, defined by its receptive field (green cylinders in Fig. 4.1), with a predefined weight matrix. The weight matrix encodes the object view, hence an object view is represented as a pattern of active/not-active V1 features in all three channels. The matrix is determined in a training phase before running the model using the trace learning method (Sec. 4.2.4).

The object localization task implies feature-based attention to the target object (Hamker, 2005b), implemented by activating a particular object neuron in the prefrontal cortex (PFC). In the primate brain, the lateral part of the prefrontal cortex contains such cells, reacting to specific objects or categories (Ashby and Spiering, 2004), and is seen as a potential source for top-down attention (Beck and Kastner, 2009; Buschman and Miller, 2007). The object neurons in PFC are connected via learned weights (Sec. 4.2.4) to view-tuned neurons in HVA layer 2/3 and further to layer 4. These connections amplify all view-tuned cells encoding the target, leading to an increased neuronal activity of all target related neurons in both HVA layers.

A model of the frontal eye field (FEF) processes spatial information and selects the target location. The area is simulated primarily by three layers: *FEFv* indicates possible locations of the target (white blobs in Fig. 4.1), *FEFvm* indicates only the final target location (single gray blob), and *FEFm* indicates the target for an upcoming eye movement (saccade). Thus, FEFvm encodes the target location during the normal processing, whereas FEFm encodes the target only if a saccade is planned towards it. Physiologically, the FEFv, FEFvm, and FEFm represent the visual, visuomovement, and movement cell types of the FEF respectively (Bruce and Goldberg, 1985; Schall, 1991). This FEF model is based on the improved FEF model of Zirnsak et al. (2011a), which represents an earlier work of the author and has been developed to enhance the FEF model's anatomical foundation.

The FEF operates as outlined in the following. FEFv is computed by taking the maximum activity over all the features in HVA layer 2/3. As target-related neurons are amplified in HVA, this computation results in a high FEFv activity at potential target locations. The signal from FEFv to FEFvm implements a soft-competition between locations by reinforcing near and suppressing remote ones. This competition favors activity on a single location, which will designate the target over time. The FEFvm projects back to HVA layer 4, forming a recurrent processing loop (HVA layer 4→HVA layer 2/3→FEFv→FEFvm→HVA layer 4). The soft-competition is continuously executed within this loop, leading to the selection of a single target location over time. The FEFm uses a similar competition to generate a saccade target. FEFm projects back to FEFvm to make the saccade target information available there too. The user can also suppress saccades by activating a fixation cell (*FEFfix*, Hasegawa et al. (2004); Hamker (2005a)) which suppresses the FEFm activity. Otherwise, a single area of activation will emerge in the FEFm over time. If this activity reaches a threshold, a saccade will be triggered towards this target location, which designates the final result of the localization process.

4.2.2. Mechanisms at neuronal level

In this section, we will describe our model at the neuronal level, i.e. its neuron types, neuronal attention mechanisms, and connection types. All neurons are modeled as a temporal-mean rate coded neuron whose response describes the temporal-averaged spike frequency of a cortical neuron (Gerstner and Kistler, 2002, Chap. 1.5). We utilize two different kinds of activation functions, the commonly-known weighted sum of excitatory and inhibitory connections (Eq. 4.1), and divisive normalization (Eq. 4.2). The former is used in the FEF, the latter in HVA by the microcircuit model of attention (Chap. 3).

The microcircuit model of attention simulates neurophysiological attention effects by a few distinct neuronal mechanisms: amplification and suppression of neuronal responses, spatial pooling in layer 2/3, and divisive normalization. Divisive normalization (Carandini and Heeger, 2012) is widely found in vertebrates, for example it accounts for the response function of neurons in the primary visual cortex to increasing stimulus contrasts (Albrecht and Hamilton, 1982). Divisive normalization uses a logarithmic, sigmoid activation function, and is combined in the microcircuit with suppression and amplification mechanisms (Eq. 4.2). Suppression has a divisive influence, thus it proportionally scales down the response. Amplification is combined with divisive normalization by a proportional scaling of the excitation E, which resembles the proportional increase of activity in cortical neurons induced by attention (Chap. 3).

$$R = E - S \tag{4.1}$$

$$R = \frac{E \cdot (1 + A)}{\sigma + E \cdot (1 + A + S)} \tag{4.2}$$

Whereby R describes the firing rate of a single neuron and σ the gain factor. E, A, S denote the sum of excitatory, amplifying, and suppressive connections towards this neuron respectively.

The system-level model contains three connection types to transport the excitatory, amplifying, and suppressive signals (Fig. 4.1):

Excitatory connection (E): a connection that has a positive weight and drives the neuron (Eq. 4.1, 4.2).

Amplifying connection (A): a connection that has a positive weight and enhances multiplicatively the neuronal activity (Eq. 4.2).

Suppressive connection (S): a connection that decreases the response. It can be implemented via a negative weight (Eq. 4.1), or via a positive weight which increases the denominator in the divisive normalization function (Eq. 4.2).

4.2.3. Attention as a holistic, cognitive control of the visual system

We see attention as a cognitive, holistic control of the visual system that tunes the whole visual processing for the task at hand. This view results from the following findings. Miller and Buschman (2013) illustrate that attentional processing spans a top-down network originating from the lateral prefrontal cortex (PFC). The PFC encodes task information (Sakai, 2008), thus we reason that this network influences neuronal activity for the current task. The network targets the ventral as well as the dorsal stream, so it affects neuronal activity in the whole visual cortex. We denote this effect as holistic. At neurophysiological level, attention has the effect to amplify the response of neurons encoding the attended stimulus, and to suppress the response of neurons encoding unattended stimuli (Chap. 3).

The attended stimulus is per definition task-relevant, hence attention modulates neuronal activity for the current task by increasing the response to task-relevant stimuli and by suppressing irrelevant ones. We combine these findings together and reason that attention is a cognitive, holistic control, modulating neuronal activity for the current task.

Our proposal is in line with previous views of attention. Carrasco (2011) defines attention as the concept to select task-relevant information among the vast amount of sensory data. Our view is a generalization of this concept, as the task control implies also to select task-relevant information. Hamker (2005a,b) see attention as an emergent result from underlying neuronal dynamics. Hamker (2005b) proposes that these neuronal dynamics guide vision. We see attention and the neuronal dynamics as two sides of the same process. Attention represents a pure psychological concept which certainly emerges from underlying neuronal dynamics. On the other hand, the neuronal dynamics realize the current task instruction which include the deployment of attention, thus the neuronal dynamics depend on attention too. Due to this strong entanglement, we see the psychological concept and the neuronal dynamics as two aspects of the same process.

Our model implements the holistic control network via top-down signals from PFC to HVA and further to the FEF. The connections from PFC to HVA layer 2/3, and downwards to layer 4, transport feature-based attention. In the object localization task, feature-based attention is deployed towards the target object (Hamker, 2005b,a). We simulate this by activating the target object cell in PFC, which triggers a signal to both HVA layers via the connections PFC→HVA layer 2/3→layer 4. This signal amplifies all HVA neurons encoding a view of the target through the amplification mechanism. Hence, this kind of amplification is denoted as *feature-based amplification mechanism*. The mechanism is combined with a *feature-based suppression mechanism*, implemented via inhibitory connections within HVA, to decrease the response of HVA neurons encoding other views. This suppresses neuronal noise, originating from distractors or background clutter. Both mechanisms are based on the microcircuit model of attention (Chap. 3).

Spatial attention emerges in the model from the recurrent processing loop between HVA and FEF. Spatial attention has the effect, via the connection from FEFvm to HVA layer 4, to amplify neurons at the attended location (*spatial amplification mechanism*), and to inhibit neurons at unattended locations (*spatial suppression mechanism*). As the target location is attended in visual search, these mechanisms enhance the target location and inhibit others. Thus, they focus neuronal activity on the target location. The first mechanism, the amplification, originates from the microcircuit model. The second mechanism, the suppression, is newly introduced to suppress the location of distractors: With real-world objects, a distractor can elicit a weak response in target encoding neurons if distractor and target object are similar enough. Thus, the target neurons have also activity at the distractor location. As those target neurons are also amplified due to feature-based attention, they must be suppressed to avoid incorrect behavior. As these neurons encode the same feature as the target, they cannot be suppressed by feature-based suppression from the target. Hence, they must be spatially suppressed, which is possible based on the information stored in the FEF. This leads to our proposal of a spatial suppression mechanism. Besides this function, the mechanism facilitates a segmentation of the target from distractors or background clutter because it inhibits those image regions. Finally, spatial attention makes the spatial information in the FEF available to HVA, and the feature-based information in HVA available to the FEF. This information exchange allows the localization and recognition of objects in parallel, which is one of the advantages of the holistic models (Chap. 2.4.3).

To summarize, the model contains the following attention mechanisms:

1a) Feature-based amplification Enhances neurons encoding an attended feature, e.g. a view, via the connection PFC→HVA layer 2/3→layer 4.

1b) Spatial amplification Enhances neurons encoding an attended location, via the connection FEFvm→HVA layer 4.

2) Spatial pooling Pools responses within the receptive field of a HVA layer 2/3 neuron.

3) Divisive normalization Normalizes the responses of HVA layer 4 and layer 2/3 neurons.

4a) Feature-based suppression Suppression between neurons encoding different features, via the connection HVA layer 2/3→layer 4. The mechanism suppresses unattended features.

4b) Spatial suppression Suppression between neurons encoding different locations, via the connection FEFvm→HVA layer 4. The mechanism suppresses unattended locations.

4c) Surround suppression Suppresses neurons encoding locations in the surround, via the connection HVA layer 2/3→layer 4. This mechanism is irrelevant for object localization, but is part of the microcircuit model (Chap. 3.3.4)

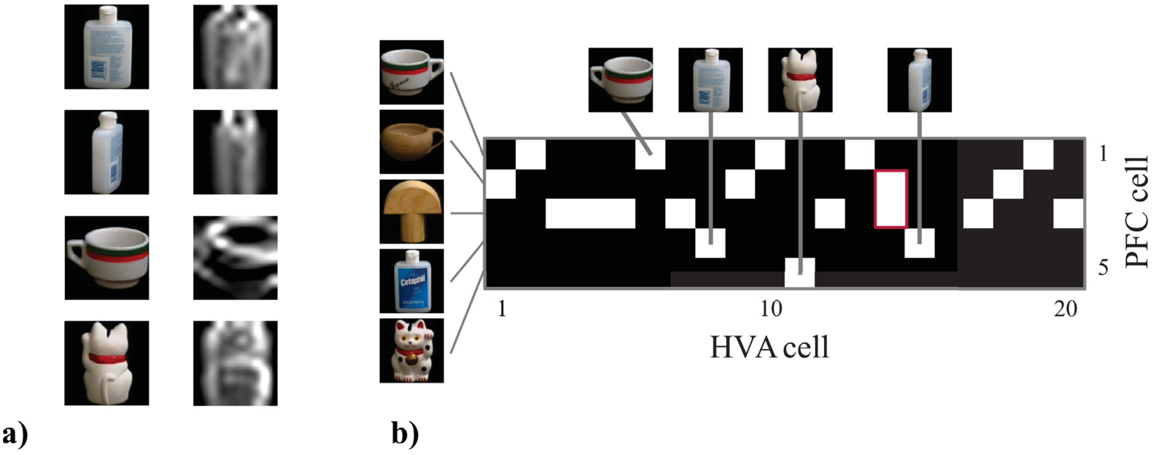


Figure 4.2.: **a)** Encoding of objects via learned view-tuned cells. For each object view (left), the weights V1→HVA of one exemplary HVA cell (right) are illustrated as the maximum over all V1 features and channels. Brightness denotes weight strength. **b)** The mapping from object cells in PFC to 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 that is incorrectly associated to multiple objects.

4.2.4. 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 (Sec. 4.4.1). 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, 1991; 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 1 000 000 presented images or 5 000 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. 4.2a). 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. 4.3). 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).

$$\begin{aligned} \tau_w \frac{\partial w_{i',i}^{\text{V1-HVA}}}{\partial t} &= \left(r_{i'}^{\text{V1}} - \theta^{\text{V1}}\right)_t \cdot \left(r_i^{\text{HVA}} - \theta^{\text{HVA}}\right)_{t-1}^{+} \\ &\quad -\alpha_w \cdot w_{i',i} \cdot \left(r_i^{\text{HVA}} - \theta^{\text{HVA}}\right)_{t-1}^{2} \\ \text{with: } \alpha_w &= \begin{cases} \alpha_{w+} & w_{i',i}^{\text{V1-HVA}} \geq 0 \\ \alpha_{w-} & w_{i',i}^{\text{V1-HVA}} < 0 \end{cases} \end{aligned} \tag{4.3}$$

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 $\operatorname{argmax}(x, 0)$. The term $\theta^{\text{V1}} = \bar{r}^{\text{V1}}$ is the mean activation of the V1 neurons, 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. 4.2b). 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. 4.3). Only such a cell has learned the presented object during the trace learning as Eq. 4.3 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. 4.2b), we keep only a single association to ensure a top-down signal specifically targeting a single object. Typically, such a cell reacts to many training 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 for this object if the scene contains both objects A and B. The task implies to amplify all HVA cells belonging to object B, but as these cells react primarily to object A, they will react more at the location of object A than of B, and the model would incorrectly select object A.

4.2.5. 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. $\boldsymbol{r}_{\boldsymbol{d},\boldsymbol{i},\boldsymbol{x}}^{\text{V1}}$). 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 $\boldsymbol{'}$ (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 $w_{x,x'}^{\text{area1-area2}}$ with the current postsynaptic neuron x and the presynaptic 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 $[x]^+$ defines a half-rectification of x:
  $[x]^+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$
- The function $[x]_0^1$ constrains x to a range between 0 and 1:
  $[x]_0^1 = \begin{cases} 1 & x \geq 1 \\ x & 0 \leq x \leq 1 \\ 0 & x < 0 \end{cases}$
- The function $x' \in \text{RF}\,(x, area)$ returns all presynaptic neuron indices x' in $area$ that are in the receptive field of the current postsynaptic neuron x.
- The response of area FEFvm is always used as a mean among its features. Thus we define: $\bar{r}_x^{\text{FEFvm}} = \frac{1}{\#i} \sum\limits_{i'} r_{i',x}^{\text{FEFvm}}$
- The function $\mathfrak{g}$ represents a two-dimensional Gaussian function, whereby x denotes the argument, a the amplitude, and σ the standard deviation. The Gaussian is always utilized centered, thus the parameter $x' = [0, 0]$. The envelope is typically chosen as $3\sigma_1 \times 3\sigma_2$ as a compromise between calculation speed and sampling precision.

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

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 that 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. 4.4), whereby $\sigma_{\text{L/M-c}}$ denotes the standard deviation of the center Gaussian and $\sigma_{\text{L/M-s}}$ of the surround Gaussian. The center signal r^c is convolved with the positive part of the DoG and the surround signal r^s with the negative part (Eq. 4.7).

$$\begin{aligned} DoG(\sigma_{\text{L/M-c}}, \sigma_{\text{L/M-s}})_x &= \mathfrak{g}(x, a_{\text{L/M-c}}, \sigma_{\text{L/M-c}}) - \mathfrak{g}(x, a_{\text{L/M-s}}, \sigma_{\text{L/M-s}}) && (4.4) \\ \text{with: } \sigma_{\text{L/M-c}} &< \sigma_{\text{L/M-s}} , && \\ a_{\text{L/M-c}} &= \tfrac{1}{2\pi\sigma^2_{\text{L/M-c}}} , \quad a_{\text{L/M-s}} = \tfrac{1}{2\pi\sigma^2_{\text{L/M-s}}} && \\ DoG^c &= a_1 \, [DoG]^+ && (4.5) \\ DoG^s &= a_2 \, [-DoG]^+ && (4.6) \\ r &= r^c * DoG^c - r^s * DoG^s && (4.7) \end{aligned}$$

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

Four different subtypes of L-M color-opponency cells were modeled (Eq. 4.8 - 4.11), 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^+.

$$\begin{aligned} r^{\text{L+M-}} &= r^{\text{L}} * DoG^c - r^{\text{M}} * DoG^s && (4.8)\\ r^{\text{M+L-}} &= r^{\text{M}} * DoG^c - r^{\text{L}} * DoG^s && (4.9)\\ r^{\text{L-M+}} &= -r^{\text{L}} * DoG^c + r^{\text{M}} * DoG^s && (4.10)\\ r^{\text{M-L+}} &= -r^{\text{M}} * DoG^c + r^{\text{L}} * DoG^s && (4.11) \end{aligned}$$

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 $\sigma_{\text{L/M-c}} = 1.5$ and $\sigma_{\text{L/M-s}} = 6$ pixels. Therefore, the size of the surround envelope is 19 pixels, corresponding to $3\,\sigma_{\text{L/M-s}}$ 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 pixels.

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. 4.12 - 4.15). The receptive field size is similar to the size of the L-M cells (Wiesel and Hubel, 1966), thus we choose identically $\sigma_{\text{S/LM}} = 6$.

$$\begin{aligned} r^{\text{LM}} &= (r^{\text{L}} + r^{\text{M}})/2 && (4.12)\\ G_{\text{S/LM}} &= \mathfrak{g}(x', a, \sigma_{\text{S/LM}}) && (4.13)\\ r^{\text{S/LM}} &= r^{\text{S}} * G_{\text{S/LM}} - r^{\text{LM}} * G_{\text{S/LM}} && (4.14)\\ r^{\text{LM/S}} &= r^{\text{LM}} * G_{\text{S/LM}} - r^{\text{S}} * G_{\text{S/LM}} && (4.15) \end{aligned}$$

Whereby the factor a normalizes $G_{\text{S/LM}}$ 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. 4.16).

$$r^{\text{GRAY}} = 0.2989\, r^{\text{R}} + 0.5870\, r^{\text{G}} + 0.1140\, r^{\text{B}} \tag{4.16}$$

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, 2005b) between L-active cells (L^+M^-, M^-L^+; Eq. 4.18) and M-active cells (M^+L^-, L^-M^+; Eq. 4.19). Each cell prefers a certain feature, namely a particular activity of L or M cells. This preference is modeled by Gaussian tuning functions (Eq. 4.17) 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. 4.20 and 4.21).

$$\begin{aligned}
H(r, \mu) &= \exp\left(-\frac{([r]^+ - \mu)^2}{2\sigma^2}\right) && (4.17)\\
r^{\text{V1S}}_{d=1,\, i=[1,4],\, x} &= H(v^{\text{Lgn}} \max\{r^{\text{L+M-}}_x,\, r^{\text{M-L+}}_x\}, \mu_i) && (4.18)\\
r^{\text{V1S}}_{d=1,\, i=[5,8],\, x} &= H(v^{\text{Lgn}} \max\{r^{\text{M+L-}}_x,\, r^{\text{L-M+}}_x\}, \mu_i) && (4.19)\\
r^{\text{V1S}}_{d=2,\, i=[1,4],\, x} &= H(v^{\text{Lgn}}\, r^{\text{S/LM}}_x, \mu_i) && (4.20)\\
r^{\text{V1S}}_{d=2,\, i=[5,8],\, x} &= H(v^{\text{Lgn}}\, r^{\text{LM/S}}_x, \mu_i) && (4.21)
\end{aligned}$$

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. 4.22). The receptive field of such an oriented-edge cells is modeled by a discretized 2D-Gabor filter (Eq. 4.23- 4.25) as proposed by Jones and Palmer (1987).

$$\begin{aligned}
\theta_i &= \{0\pi,\ 0.25\pi,\ ...\ 1.75\pi\} && (4.22)\\
X_{1,i} &= x_1 \cos\theta_i + x_2 \sin\theta_i && (4.23)\\
X_{2,i} &= -x_1 \sin\theta_i + x_2 \cos\theta_i && (4.24)\\
G_{i,x} &= A \cdot exp\left(-\left(\frac{X_{1,i}^2}{2\sigma_1^2} + \frac{X_{2,i}^2}{2\sigma_2^2}\right)\right) \cdot \cos(2\pi f X_{1,i} + \psi) && (4.25)\\
r^{\text{V1S}}_{d=3,\,i,\,x} &= r^{\text{GRAY}} * G_i && (4.26)
\end{aligned}$$

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 as in 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. 4.28) In addition, response differences are enhanced via a non-linearity (Eq. 4.29). 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 smoothes out the response at the borders and avoids the problem. As kernel, we use a Lanczos3 kernel (Eq. 4.27) as it meets the requirements and is an often-used standard approach for decreasing resolutions (Turkowski and Gabriel, 1990).

$$\begin{aligned}
K_x &= \begin{cases} 1 & x' = 0 \\ \dfrac{a \sin(\pi\, x')\, \sin(\pi\, x'/a)}{\pi^2\, x'^2} & 0 < |x'| < a \\ 0 & x' \geq a \end{cases} && (4.27)\\
\text{with} &: \quad a = 3, \quad x' = x/10
\end{aligned}$$

$$R_{d,i,x} = r^{\text{V1S}}_{d,i} * K * K^T \tag{4.28}$$

$$r^{\text{V1C}}_{d,i,x} = R^{p_{V1C}} \tag{4.29}$$

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) 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., 2007a; 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. 4.3). 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 chapter 3 for its anatomical and neurophysiological background. The microcircuit model is slightly differently parameterized than in chapter 3 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 ($A^{\text{FEAT-L2}}$) 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 (Sec. 4.4.3). As second change, the system-level model has a larger number of cells, and thus the weighted sums operate in many connections over more presynaptic cells. This is a problem when these presynaptic cells are weakly active, for example from noise, as the weak activity would sum up to a significant signal due to the sheer number of cells. To avoid this problem, we use higher non-linearities in the weighted sum to diminish the influence of weak presynaptic cells, and two non-linearities for a finer parameter tuning. A typically example for the higher non-linearities is the feature-based suppression (S^{FEAT}): The microcircuit model uses the non-linearity function $f(x) = x^2$ for 33 cells in the feature space (Fig. 3a in Chap. 3.3.1), while the system-level model uses a composition of the functions $f(x) = x^3$ and $f(x) = x^5$ for 354 cells. As third 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

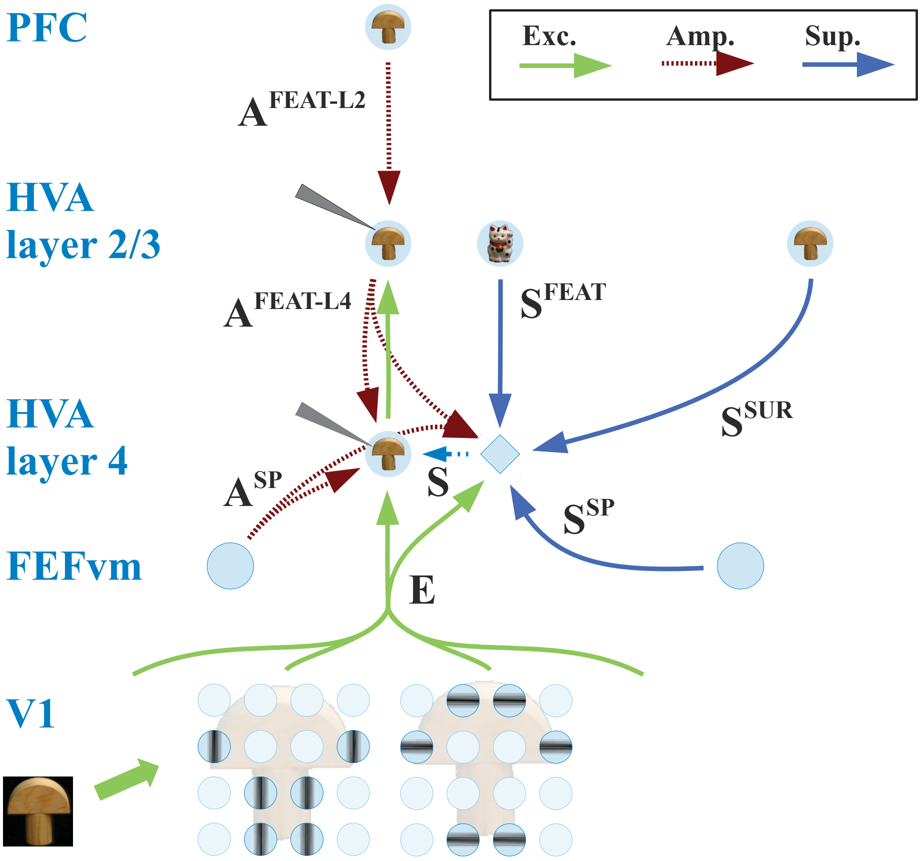


Figure 4.3.: Detailed illustration of the higher visual area (HVA), when localizing an object. The stimulus, here a “wooden mushroom”, excites a specific spatial response pattern in each feature plane in V1, exemplary shown for two planes encoding vertical and horizontal edges. Excitation (E) of HVA layer 4 is calculated from this pattern via a weighted sum. The remaining of the figure shows the connectivity and influences on a single cell encoding the object in HVA layer 4 and 2/3, indicated by the electrode symbols. The layer 4 cell receives feedforward excitation from V1 (E), feature-based amplification ($A^{\text{FEAT-L4}}$) from layer 2/3, spatial amplification from FEFvm (A^{SP}), 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 ($A^{\text{FEAT-L2}}$).

required part of the target selection process (Sec. 4.2.1 and Sec. 4.3.2) 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 $h = 1\,ms$, and are constrained to $[0, 1]$ by the function $[x]^+$.

$$\tau^{\text{HVA4}} \frac{\partial r^{\text{HVA4}}_{d,i,x}}{\partial t} = -r^{\text{HVA4}}_{d,i,x} + g^{\text{HVA4}} \cdot \frac{E_{d,i,x} \cdot A_{d,i,x}}{\sigma + S_{d,i,x}} \tag{4.30}$$

$$S_{d,i,x} = E_{d,i,x} \cdot \left(A_{d,i,x} + S^{\text{FEAT}}_{d,i,x} + S^{\text{SP}}_{x} + S^{\text{SUR}}_{d,i,x}\right) \tag{4.31}$$

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, whereas it encodes target and mask in the object substitution masking scenario (Chap. 5). 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 ($d = \{1, 2, 3\}$), whereas 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. 4.32), or via pooling of V1 features (Eq. 4.33). Both variations can be scaled via $v^{\text{V1-HVA4}} = 1$ and a non-linearity parameter $p^{\text{E}} = 1$. The connectivity weights $w^{\text{V1-HVA4}}$ are either provided by an external learning procedure or modeled by a 2D-Gaussian (Eq. 4.34).

$$E_{d,i,x} = \left[v^{\text{V1-HVA4}} \cdot \left[\sum_{d',i',x' \in \text{RF}(x,V1)} \left(w^{\text{V1-HVA4}}_{i,\,d',i',x'}\, r^{\text{V1C}}_{d',i',x'} \right) \right]_0^1 \right]^{p_E} \tag{4.32}$$

$$E_{d,i,x} = \left[v^{\text{V1-HVA4}} \cdot \left[\max_{x' \in \text{RF}(x,V1)} \left(w^{\text{V1-HVA4}}_{x'}\, r^{\text{V1C}}_{d,i,x'} \right) \right]_0^1 \right]^{p_E} \tag{4.33}$$

$$w_{x'}^{\text{V1-HVA4}} \quad = \quad \mathfrak{g}(x', 1, [8.3, 8.3]) \tag{4.34}$$

A cell receives spatial amplification from the same location in FEFvm (Eq. 4.36), and feature-based amplification from the same feature in HVA layer 2/3 (Eq. 4.37). 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. 4.38), 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. 4.35) as multiple studies show an additive influence of both attention forms (Saenz et al., 2002; Treue and Trujillo, 1999).

$$\begin{aligned}
A_{d,i,x} &= 1 + A_{d,i,x}^{\text{SP}} + A_{d,i,x}^{\text{FEAT-L4}} && (4.35)\\
A_{d,i,x}^{\text{SP}} &= v^{\text{FEFvm-HVA4}}\, \overline{r}_x^{\text{FEFvm}} && (4.36)\\
A_{d,i,x}^{\text{FEAT-L4}} &= v^{\text{HVA2-HVA4}} \left(\max_{x' \in \text{RF}(x, HVA2)} w_{x'}^{\text{HVA2-HVA4}}\, r_{d,i,x'}^{\text{HVA2}} \right)^{p_{HVA2-HVA4}} && (4.37)\\
w_{x'}^{\text{HVA2-HVA4}} &= \mathfrak{g}(x', 1, [0.6, 0.6]) && (4.38)
\end{aligned}$$

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. 4.39). 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. 4.40).

If HVA encodes simple features:

$$w_{d,i,i'}^{\text{FEAT}} = \begin{cases} (|i-i'|/7)^2 & d = 1, 2 \\ (|i-i'|/3)^2 & d = 3, |i-i'| \leq 3 \\ 1 - (|i-i'| - 4)/3)^2 & d = 3, |i-i'| > 3 \end{cases} \tag{4.39}$$

If HVA encodes object views:

$$w^{\text{FEAT}}_{d,i,i'} = \begin{cases} 0 & \text{Views cells } i \text{ and } i' \text{ belong to the same object.} \\ 1/\#L_k & \text{Else.} \end{cases} \tag{4.40}$$

The strength in the later case is normalized with the number of view cells encoding a particular object (Eq. 4.40). View cells encoding the same object k are typically simultaneously active, thus there exist $\#L_k$ presynaptic cells i' which simultaneously suppress a postsynaptic cell i. A normalization with $1/\#L_k$ ensures that the postsynaptic cell i will receive the same amount of suppression, independently of the number of presynaptic cells $\#L_k$.

The feature-based suppression (Eq. 4.43) 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 (Chap. 3) which proposes feature-based suppression only within a local region, similar in size as the receptive field. The earlier model of Hamker (2005a) 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. 4.42) 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. 4.38) as the feature-based amplification.

Non-linearities are implemented via power-functions ($f(x) = x^p$) 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.

$$S^{\text{FEAT}}_{d,i,x} = \left[v^{\text{FEAT-1}} \cdot \left[\sum_{i'} w^{\text{FEAT}}_{d,i,i'} \cdot (v^{\text{FEAT-2}}\, B_{d,i',x})^{pFEAT-2} \right]_0^1 \right]^{pFEAT-1} \tag{4.41}$$

If HVA encodes simple features:

$$B_{d,i,x} = \max_{x' \in \text{RF}(x,HVA2)} \left(w_{x'}^{\text{HVA2-HVA4}} \, r_{d,i,x'}^{\text{HVA2}} \right) \tag{4.42}$$

If HVA encodes object views:

$$B_{d,i,x} = \max_{x'} \left(r_{d,i,x'}^{\text{HVA2}} \right) \tag{4.43}$$

Spatial suppression is received from other retinotopic locations in the FEFvm (Eq. 4.44). Its parameters control again separately the non-linear influence of presynaptic 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. 4.45) 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. Next, we found out 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 (Sec. 4.2.3), and we realize this via such inhibitory connections.

$$S_x^{\text{SP}} = \left[v^{\text{SP-1}} \cdot \sum_{x' \in \text{RF}(x,FEFvm)} w_{x'}^{\text{SP}} \cdot \left(v^{\text{SP-2}} \, \bar{r}_{x'}^{\text{FEFvm}} \right)^{pSP-2} \right]^{pSP-1} \tag{4.44}$$

$$w_{x'}^{\text{SP}} = \left[1 - 2 \left(\mathfrak{g}(x', 1, [3,4])^{0.125} \right) \right]^{+} \tag{4.45}$$

Surround suppression is received from layer 2/3 neurons encoding the same feature at surround locations (Eq. 4.47). 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. 4.46).

$$w_{x'}^{\text{SUR}} = \frac{K}{|K|}, \text{ with: } K = [\mathfrak{g}(x', 1, [3,3]) - \mathfrak{g}(x', 2, [1.5, 1.5])]^{+} \tag{4.46}$$

$$S_{d,i,x}^{\text{SUR}} = \left[v^{\text{SUR-1}} \cdot \sum_{x' \in \text{RF}(x, HVA2)} w_{x'}^{\text{SUR}} \cdot \left(v^{\text{SUR-2}} \, r_{d,i,x'}^{\text{HVA2}} \right)^{pSUR-2} \right]^{pSUR-1} \tag{4.47}$$

Higher visual area - layer 2/3

HVA layer 2/3 pools spatially layer 4 responses (Eq. 4.48, 4.49), whereby the pooling is executed for each feature and channel separately.

$$\tau^{\text{HVA2}} \frac{\partial r_{d,i,x}^{\text{HVA2}}}{\partial t} = -r_{d,i,x}^{\text{HVA2}} + g^{\text{HVA2}} \cdot \frac{E_{d,i,x} \cdot (1 + A_{d,i,x}^{\text{FEAT-L2}})}{\sigma + S_{d,i,x}} \tag{4.48}$$

$$S_{d,i,x} = E_{d,i,x} \cdot (1 + A_{d,i,x}^{\text{FEAT-L2}}) \tag{4.49}$$

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 $g^{\text{HVA2}} = 1.69$ 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. 4.50), whereby $v^{\text{HVA4}} = 1$ controls its influence. The pooling is implemented via a soft-max (Chap. 3), whereby $p_1 = 4$ and $p_2 = 0.25$ parameterize the involved non-linearities. The connectivity is implemented via a 2D-Gaussian, forming a receptive field with strongest excitation in the center and weaker ones at the borders (like found for example by Heuer and Britten (2002) for area MT). A Gaussian-modulated pooling has been used also in previous models, e.g. Hamker (2005b,a); Hamker and Zirnsak (2006).

$$E_{d,i,x} = \left(v^{\text{HVA4}} \cdot \sum_{x' \in \text{RF}(x, HVA4-HVA2)} w_{x'}^{\text{HVA4-HVA2}} \left(r_{d,i,x'}^{\text{HVA4}} \right)^{p_1} \right)^{p_2} \tag{4.50}$$

$$w_{x'}^{\text{HVA4-HVA2}} = \mathfrak{g}(x', 1, [1,1]) \tag{4.51}$$

Feature-based amplification is received from the prefrontal cortex (PFC, Eq. 4.52), 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 (Beck and Kastner, 2009; Buschman and Miller, 2007; Fuster, 2000; Tomita et al., 1999). The signal is part of the attentional processing network in the cortex (Chap. 2.2). 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).

$$A_{d,i,x}^{\text{FEAT-L2}} = v^{\text{PFC-HVA2}} \cdot \sum_{i'} w_{d,i,i'}^{\text{PFC-HVA2}} \, r_{d,i'}^{\text{PFC}} \tag{4.52}$$

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 (Chap. 2.2).

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. 4.2.4). We simulate one cell for each object i: r_i^{PFC}. 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: $r_{d,i}^{\text{PFC}}$. This configuration is used in object substitution masking (Chap. 5), and might also be appropriate 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. (2011a). 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 (Bruce and Goldberg, 1985; 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, 2005b). This representation is often denoted as saliency map (Itti and Koch, 2001).

The FEFv is excited from HVA layer 2/3 (Eq. 4.56) and from V1 (Eq. 4.57). 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. 4.58). 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 will need it in object substitution masking (Chap. 5), and we include it also for the generality of the model. The neurons in the FEF use as activation function a half rectification function to ensure positive responses ($[\,]^+$). This was not necessary in the divisive normalization model, as it ensure positive responses automatically.

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. 4.59) (Antonelli et al., 2014), and by a signal enhancement operation Q based on divisive normalization (Eq. 4.60). 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.

$$
\begin{aligned}
\tau^{\text{FEFv}} \frac{\partial r_x^{\text{FEFv}}}{\partial t} &= -r_x^{\text{FEFv}} + E_x && (4.53)\\
\text{with: } \tau^{\text{FEFv}} &= 10\,, \quad \text{and} \quad r_x^{\text{FEFv}} = \left[r_x^{\text{FEFv}}\right]^+ \\
E_x &= C\left(Q\left(F_x\right)\right) && (4.54)\\
F_x &= \max\left\{E_x^{\text{V1}}, E_x^{\text{HVA2}}\right\} && (4.55)\\
E_x^{\text{HVA2}} &= \max_{d',i'}\left(r_{d',i',x}^{\text{HVA2}}\right) && (4.56)\\
E_x^{\text{V1}} &= v^{\text{V1-FEFv}} \cdot \max_{d',i'}\left(\left(\sum_{x' \in \text{RF}(x,FEFv)} w_{x'}^{\text{V1-FEFv}} \left(r_{d',i',x'}^{\text{V1C}}\right)^{p_1}\right)^{p_2}\right) && (4.57)\\
w_{x'}^{\text{V1-FEFv}} &= \mathfrak{g}(x', 1, [36.6, 36.6]) && (4.58)\\
C(x) &= [x \cdot (1+c) - c]^+ && (4.59)\\
Q(x) &= x \cdot \frac{(1+\sigma)}{F^{\text{max}} + \sigma} && (4.60)
\end{aligned}
$$

Whereby E denotes the excitation, τ^{FEFv} 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 (Schall, 1991; Ray et al., 2009). Our modeled cells encode a continuous spectrum of both influences (Zirnsak et al., 2011a). The visual information is transmitted via inputs from FEFv (E^{FEFv} in Eq. 4.63), and the saccadic information via inputs from FEFm (r^{FEFm} in Eq. 4.61). Their influences are proportionally weighted to each other via $v_i^{\text{FEFv-FEFvm}}$ and $1 - v_i^{\text{FEFv-FEFvm}}$ (Eq. 4.61).

$$
\begin{aligned}
\tau^{\text{FEFvm}} \frac{\partial r_{i,x}^{\text{FEFvm}}}{\partial t} &= -r_{i,x}^{\text{FEFvm}} + v_i^{\text{FEFv-FEFvm}} E_x^{\text{FEFv}} + \left(1 - v_i^{\text{FEFv-FEFvm}}\right) r_x^{\text{FEFm}} && (4.61)\\
\text{with: } \tau^{\text{FEFvm}} &= 10\,, \quad \text{and} \quad r_{i,x}^{\text{FEFvm}} = \left[r_{i,x}^{\text{FEFvm}}\right]^+
\end{aligned}
$$

The input signal from FEFv realizes a competition between locations via a local Gaussian excitation (Eq. 4.63) and a long-range suppression (Eq. 4.64). Such a long-range, spatial competition is typically employ by models of visual search, despite its neuronal mechanisms are not known. 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. 4.64). The excitation can be scaled similarly via $v^{\text{Ev}} = 0.6$ (Eq. 4.63). 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. 4.62).

$$\begin{aligned}
E_x^{\text{FEFv}} &= v^{\text{low}} \cdot E_x + (1 - v^{\text{low}}) \cdot [E_x - S_x]_0^1 && (4.62)\\
E_x &= v^{\text{Ev}} \cdot \sum_{x'} w_{x'}^{\text{Ev}}\, r_{x'}^{\text{FEFv}} && (4.63)\\
S_x &= \left(v^{\text{Sv-1}} \cdot \sum_{x'} w_{x'}^{\text{Sv}}\, r_{x'}^{\text{FEFv}} \right)^{p_{Sv-1}} && (4.64)\\
K_{x'} &= \mathfrak{g}(x', 1, [3, 4]) - v^{\text{Sv-2}} && (4.65)\\
w_{x'}^{\text{Ev}} &= [K_{x'}]^+ && (4.66)\\
w_{x'}^{\text{Sv}} &= [-K_{x'}]^+ && (4.67)
\end{aligned}$$

Frontal eye field - movement (FEFm) and fixation cells (FEFfix)

The FEFm represents 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, 2005a). FEFm (Eq. 4.68) 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. 4.69) and a long range inhibition (Eq. 4.70). 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, 2005a) 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. 4.71). 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.

$$\begin{aligned}
\tau^{\text{FEFm}} \frac{\partial r_x^{\text{FEFm}}}{\partial t} &= -r_x^{\text{FEFm}} + E_x^{\text{FEFvm}} - S^{\text{FEFvm}} - S^{\text{Fix}} && (4.68)\\
\text{with: } \tau^{\text{FEFm}} &= 10\,, \quad \text{and} \quad r_x^{\text{FEFm}} = [r_x^{\text{FEFm}}]^+ &&\\
E_x^{\text{FEFvm}} &= v^{\text{FEFvm-FEFm}}\, \overline{r}_x^{\text{FEFvm}} && (4.69)\\
S^{\text{FEFvm}} &= v^{\text{Svm}} \max_{x'} \overline{r}_{x'}^{\text{FEFvm}} && (4.70)\\
S^{\text{FEFfix}} &= v^{\text{Sfix}}\, r^{\text{FEFfix}} && (4.71)
\end{aligned}$$

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, 2005a) 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. 4.72).

$$x_c = \frac{\sum_{x'} r_{x'}^{\text{FEFm}}(t_o) \cdot x'}{\sum_{x'} r_{x'}^{\text{FEFm}}(t_o)} \qquad (4.72)$$

4.3. The role of neuronal attention mechanisms in object localization

According to our models, attention controls the visual system with a few, neuronal mechanisms, and we would like to investigate their role in the task of object localization. Our model inherits from the neurophysiologically-grounded microcircuit model of attention the neuronal mechanisms of amplification, normalization, spatial pooling, and suppression. Here, we will focus on the amplification and the suppression mechanisms as such mechanisms are also part of other top-down and holistic attention models. Both mechanisms exist as a feature-based and a spatial variant. We will concentrate on the feature-based variants, i.e. feature-based amplification and suppression, as object localization is performed by deploying feature-based attention to the target object and thus this kind of attention seems important for the task. Besides this main reason, we also leave out the spatial variants as we cannot disable them easily for the evaluation. A disabling would imply to cut the FEF→HVA connections, which would destroy the ability of the model to focus neuronal activity to a single location, and thus to localize objects.

4.3.1. Feature-based amplification represents the target

Before we shed light on the role of the amplification mechanism, we illustrate the mechanism's neuronal processing in object localization (Fig. 4.4a). The mechanism is part of the attentional control network and amplifies multiplicatively cells encoding the attended object (Sec. 4.2.3). In the object localization task, the target object is attended by its type (feature-based attention). In the model, the control network is simulated via the connections from PFC to HVA layer 2/3 and layer 4. The feature-based mechanism realizes the influence of these connections on the neurons in HVA. It amplifies all HVA neurons encoding a view of the target object.

The feature-based amplification mechanism has the role to represent the target of the current localization task in the visual system. It amplifies all neurons encoding the target, and in this way, represents the target in the system. Related to this general role, the amplification mechanism serves two purposes at the level of neuronal signals.

Firstly, the amplification differentiates the target signal from any signals in target unrelated neurons, denoted as neuronal noise. For this case, we consider the recognition of a single

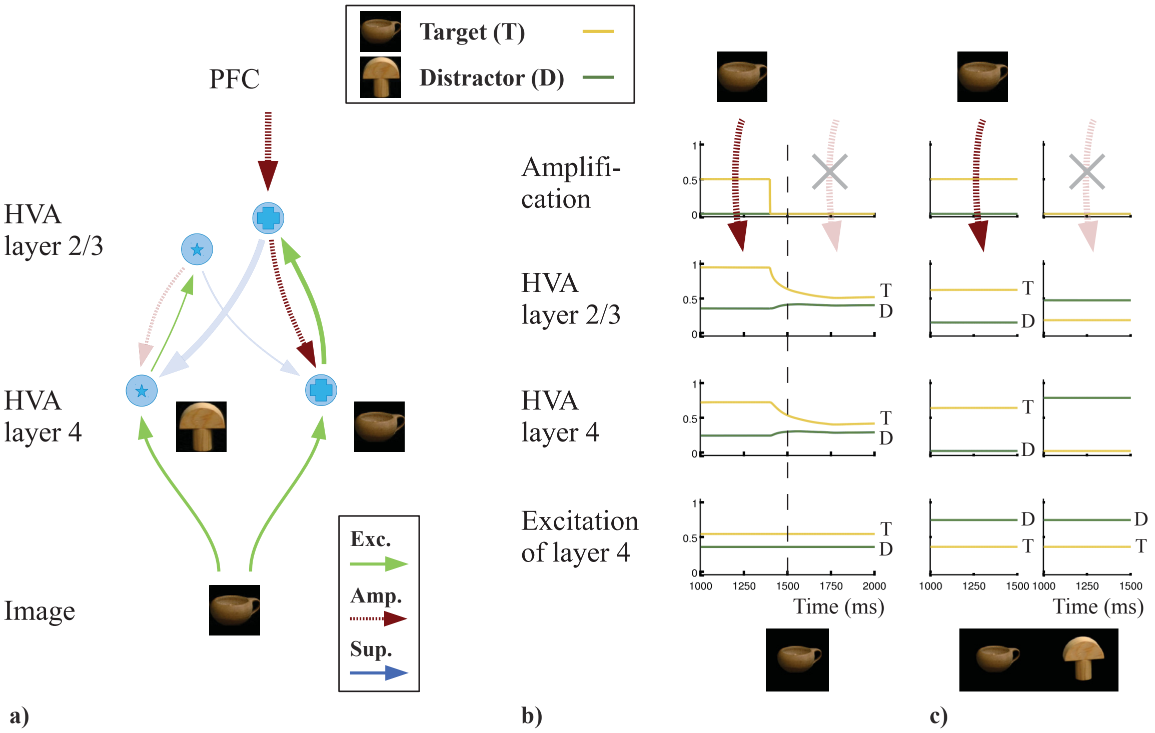


Figure 4.4.: a) The neuronal mechanism of feature-based amplification in an object localization scenario. The mechanism is mediated by amplifying connections (red arrows). The connection type (excitation, amplification, suppression; see Sec. 4.2.2) describes the influence on its postsynaptic neuron. In each layer, representative neurons encoding the target (“wooden cup”) and the distractor object (“wooden mushroom”) are shown. The objects are taken from the COIL-100 database (Nene et al., 1996), whereby their encoding has been previously learned via the trace learning procedure (Chap. 4.2.4). Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness. For clarity, unrelated connections are not illustrated in the figure. **b, c)** Effect on the neuronal responses. The feature-based amplification enhances the responses of neurons encoding the attended target (yellow, T), which increases their difference with the responses of neurons encoding a distractor (green, D). The effect is illustrated by disabling the mechanism, whereby in both (b) and (c), it is enabled at the left and disabled at the right. When enabled, the mechanism enhances the target signal to differentiate it from other distractor signals (shown in b). When the initial target response is weak (e.g. non-salient target), the amplification increases the target signal over the distractor signal (shown in c), selecting the target at neuronal level and leading to a clear neuronal code. A distractor encoding neuron can be obviously driven by the distractor stimulus (c), but as well by the target if it is similar to the distractor (b).

object, the “wooden cup” (Fig. 4.4b). Noise emerges as the stimulus does not only excite HVA layer 4 neurons encoding it, but also neurons encoding similar distractors (“wooden mushroom”). The two objects are similar due to their color (both brown objects). Attending the target amplifies the response of the target neuron, so the target signal can be better differentiated from the distractors signal. The effect occurs in HVA layer 4 as well as in layer 2/3. We illustrate the amplification mechanism by disabling it via cutting the connection from PFC to HVA layer 2/3 after 1300 ms. After deactivation, a slow decrease of the target response can be observed until it is similar to its excitation. Methodically, we keep the excitation to HVA layer 4 constant by presenting the stimulus continuously during the experiment which avoids network oscillations, and we start recording responses after 1000 ms which avoids stimulus onset effects. The strengths of amplification and suppression signals were chosen according to the values fitting at best the physiological data (standard values in Tab. 3.1, Chap. 3).

Secondly, the amplification selects the target from a target-distractor pair by increasing the target signal over the distractor signal (Fig. 4.4c). The illustrated scene is problematic for object localization as the target is less salient than the distractor, hence the excitation of the target neuron is lower than of the distractor neuron. Attention removes this difference by amplifying the target signal above the distractor signal. In combination with the suppression mechanism, the amplification leads to a clear neuronal code containing only the target signal (left in Fig. 4.4c). Based on this code, the FEF selects correctly the target. If the amplification is disabled (right), the distractor signal remains high and the target signal is decreased due to suppression. Thus, without attention, the system will focus the most salient object in the scene. Methodically, we invoke two experiments to process the same scene with and without attention to the target. One experiment was not possible due to the spatial network dynamics, which would maintain a spatial amplification of the target location even after disabling of feature-based amplification. This was irrelevant in the first case as the scene contains an object only at a single location.

In the cortex, amplifications seems to occur in all visual areas as attention effects were found in higher, but also in lower visual areas (Saenz et al., 2002). Lower visual areas encode simple features like colors, edges (V1) or corners (V2). The presence of attention effects in these areas implies that also parts of an object like particular edges or colors receive amplification. This indicates that attention indeed tunes the whole visual system from early to higher areas for recognizing an object. We implemented exemplary the amplification effect for the two layers of HVA. However, a neuronal modulation could also be realized in earlier visual areas as an amplifying signal from HVA to earlier visual cortices could be implemented in a similar way to the existing signal from PFC to HVA.

In summary, attention performs object localization by the feature-based amplification mechanism. It represents the target in the visual system by amplifying target specific neurons in all layers, which tunes the whole visual system for processing the target's features.

4.3.2. Feature-based suppression removes noise

Attention is typically combined with suppression to remove noise (Beuth et al., 2014). Noise is defined here as any signal overlaying the target object signal (signal detection theory, Swets (2014)). Noise can originate from external or internal sources. Examples for external noise are a blurring of the input image, a white-noise background, a cluttered scene, or distracting objects. Internal noise results from the fact that a neuron reacts to a stimulus independent of small changes, for example to object rotations of up to 20 degree (Logothetis et al., 1995). However, this response behavior implies that a certain stimulus excites not only the neuron encoding this stimulus, but also neurons encoding similar stimuli. The latter neuronal response is denoted as internal noise.

Feature-based suppression reduces the neuronal noise as it suppresses the neuronal responses of unattended features (Sec. 4.2.3). In the proposed model, this occurs via suppressive connections from layer 2/3 to layer 4 (Sec. 4.2.2). The amplified response of the target encoding neurons increases the suppression towards neurons encoding a distractor, decreasing the neuronal responses firstly in layer 4 and afterwards in layer 2/3 (Fig. 4.5a). Suppression is typically observed between divergent features like 90 degree rotated edges (MacEvoy et al., 2009). Here, suppression occurs between views belonging to different objects.

Evidence for the suppression mechanism comes from the phenomenon of biased competition (Desimone and Duncan, 1995; Treue and Trujillo, 1999). According to the biased-competition paradigm, competition takes place when two different stimuli are presented inside a receptive field of a neuron. In the unattended condition, both stimuli fire less than in a condition where a single stimulus is presented, indicating suppression between them. However, if attention is directed to one of the stimuli, the neuron encoding this stimulus fires stronger, whereas the neuron encoding the unattended stimulus is suppressed. The phenomenon has also been reported in an object localization setup by Chelazzi et al.

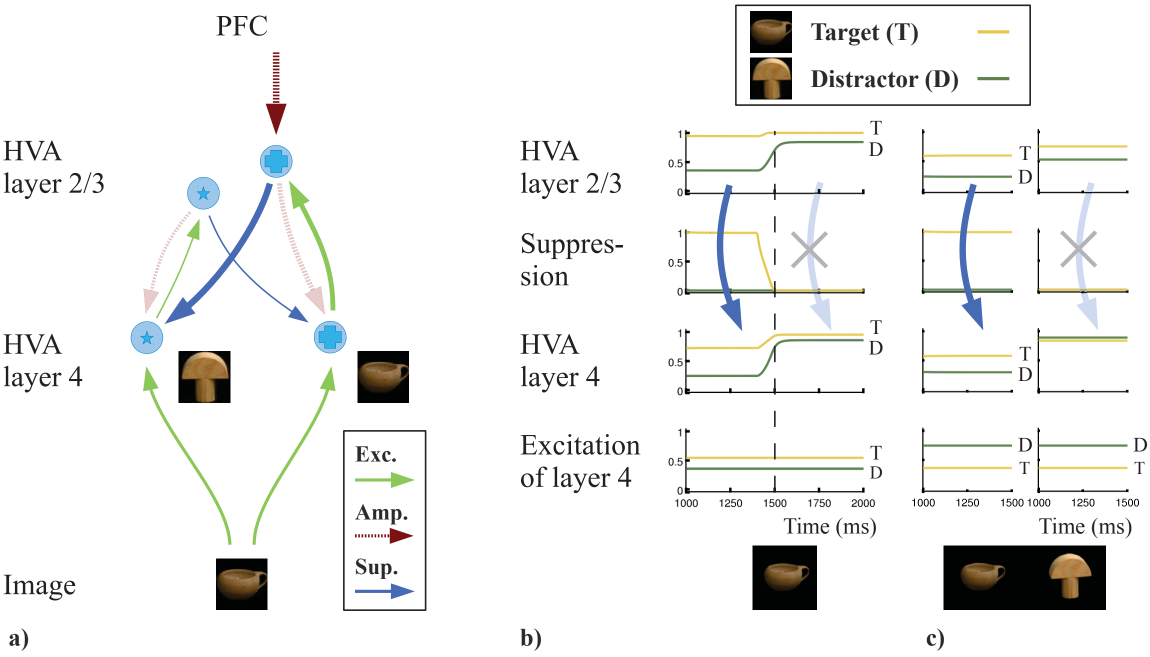


Figure 4.5.: a) The neuronal mechanism of feature-based suppression in an object localization scenario. The mechanism is mediated by suppressive connections (blue arrows). The notation of the figure is similar to Fig. 4.4. **b, c)** Effect on the neuronal responses. The suppression from layer 2/3 to layer 4 decreases strongly the response of a distractor neuron in layer 4, resulting also in a weak response of layer 2/3 neurons. The mechanism decreases neuronal noise in the system, defined as all neuronal responses unrelated to the target, like the responses of the distractor neurons. These responses can obviously result from a distractor stimulus (c), but as well from the target if it is similar to the distractor (b). In both illustrations, the suppression mechanism is enabled at the left, and disabled at the right by cutting the suppressive connection from layer 2/3 to layer 4.

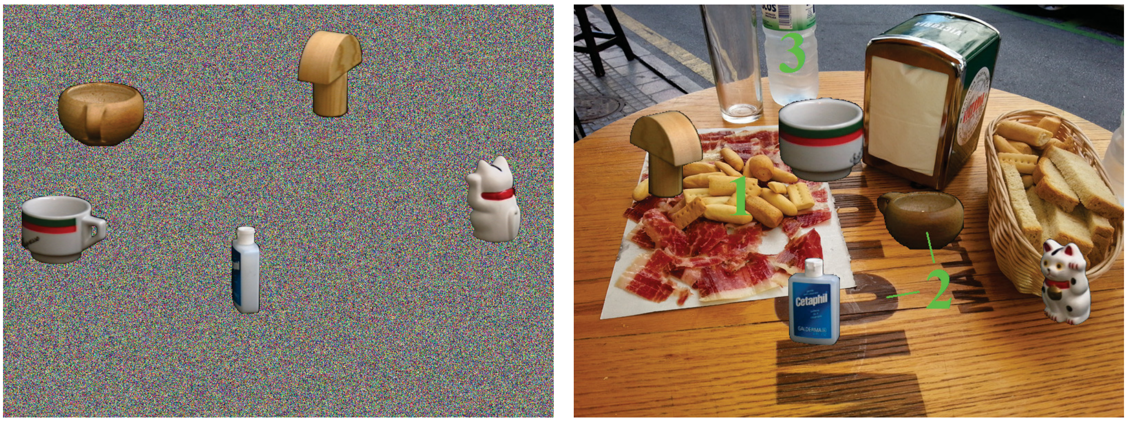
(1998), using a scene with a target and a distractor object like in Fig. 4.5c and recording area IT which is comparable to HVA. The microcircuit model underlying HVA explains this data set with the proposed suppression mechanism (Chap. 3).

We now investigate the suppression of internal and external noise, and utilize for this purpose the same scenes as in last section: a scene with a single target object (Fig. 4.5b), and a scene with a target and a distractor (Fig. 4.5c). The first scene allows to examine the internal noise, as the presented target object excites neurons encoding this target, but also excites moderately neurons encoding the distractor (Fig. 4.5b). The experimental conditions are identical as in the previous section: we use a constant input, and the suppression strength parameters were set according to the physiological data (Chap. 3). The target object was attended which amplifies the response of the target neurons in HVA (yellow). This amplified response in HVA layer 2/3 increases the suppression to the neuron encoding the distractor (green), resulting in a low response in HVA layer 4. To show the effect of the suppression mechanism, it was disabled at 1500 ms by decreasing the weight of the suppressive connection to zero within 200 ms. Without suppression, the distractor neuron reacts strongly. Therefore, the suppression reduces internal noise like the response of distractor neurons towards a target stimulus.

The mechanism is similarly useful to reduce external noise, for example originating from a distractor (Fig. 4.5c). The scene was processed with and without suppression. In the experiment with suppression enabled (shown on the left), the responses of the distractor neurons are suppressed. Without suppression (right), the responses of distractor neurons in layer 4 are very high, similar to the one of target neurons. Note, the responses of target neurons in layer 2/3 are slightly higher than of distractor neurons due to the amplification of the target. Suppression was disabled by cutting the suppressive connections from layer 2/3 to layer 4. The experimental setup is again identical as in the last section, except that spatial suppression is deactivated as we would like to examine only feature-based suppression without any interference from other suppression mechanisms. The weaker responses with suppression show that the mechanism reduces also external noise.

Importantly, the target was correctly localized in all conditions, even while feature-based suppression was disabled. To understand this fact, we consider two neurons suppressing each other. Disabling the feature-based suppression does not change which of the two neurons is the most active one (Fig. 4.5b, c). However, the localization process of the FEF depends only on the most active neurons, so the suppression will not affect the localization performance.

We conclude that the feature-based suppression decreases the neuronal noise and does not affect the localization performance. The same mechanism accounts for the suppression of neuronal activity in the biased competition phenomena. Hence, our findings shed light on the role of this neurophysiological effect.



(a) (b)

Figure 4.6.: Exemplary benchmark scenes with colored white-noise background **(a)** and with real-world background **(b)**. Green numbers mark properties of the background (see main text).

4.4. Results

We will now evaluate the performance of the full system, as well as the performance impact of the attention mechanisms to clarify their role in object localization and to illustrate the benefits of attention. Here we will focus again on the common mechanisms for enhancing the discrimination among objects, i.e. the feature-based amplification and the feature-based suppression. We will benchmark each of these mechanisms individually by disabling it and by measuring the drop in the localization accuracy.

4.4.1. Evaluation methods

To benchmark the system, we used whole scenes composed of three differently cluttered backgrounds and arbitrary objects from the COIL-100 database (Columbia Object Image Library, Nene et al. (1996)). Fig. 4.6 shows two examples. The COIL-100 database contains 100 objects (Fig. B.1 in the Appendix B) recorded from all sides in 5° steps before a dark-gray background at 128x128 pixel resolution, resulting in 7200 images in total. The database is very suitable for learning rotation invariance due to its high number of objects and rotations, and thus widely used (Einhäuser et al., 2005; Kietzmann et al., 2009; Lessmann and Würtz, 2014; Mobahi et al., 2009; Pinto et al., 2008). As the images represent a continuously rotation over time, it is also utilized for learning based on temporal continuity

(Einhäuser et al., 2005; Mobahi et al., 2009). We also employ both approaches, and thus use the database for the same reasons.

The recognition accuracies on the COIL-100 database are quite high for recognizing single objects in a scene (Lessmann and Würtz, 2014). Specialized computer vision systems archive for example a 100% recognition accuracy for a splitting of training to test set of 50%:50% (Lessmann and Würtz, 2014) and 70% for a 5%:95% split (Lessmann and Würtz, 2014). Other reported even higher performances (Lessmann and Würtz (2014) on the data of Linde and Lindeberg (2012), Westphal and Würtz (2009)). However, these systems lack biological plausibility, for example Lessmann and Würtz (2014) has no locality in their learning rule. Another difference is that all these studies recognize single objects, whereas we evaluate complete scenes in a localization task, which is much more realistic, but also more challenging. Because of these reasons, we omit a direct comparison with the computer vision systems. In biological systems, the focus is on the precise modeling of the cortex and not on the highest performance, so they typically perform slightly worse. For example, Einhäuser et al. (2005) reported 75% recognition accuracy for a 16%:84% split with 50 objects and scenes composed of single objects. Scenes with multiple objects were used by Elazary and Itti (2010) to evaluate their model in a localization task. The model belongs to the class of probabilistic, top-down saliency models, so it is biologically inspired, but not so well founded as ours or Einhäuser's model. They achieve a localization accuracy of 97%, whereby their setup uses black backgrounds and a 50%:50% split.

We created nine test sets composed of 50 cluttered scenes, downloadable via: `https://www.tu-chemnitz.de/cs/KI/supplement/Beuth2019/`.
The sets include three different backgrounds and three different subsets of objects from the COIL-100 database, resulting in nine test sets. The three object subsets consist of 5 objects, 15 arbitrary chosen objects, and all 100 objects. The test scenes were created for each number of objects separately by using only objects from the subset. The five-object set contains the objects 10-14 from the COIL-100, namely the "porcelain cup", "wooden cup", "wooden mushroom", "cleaning bottle", and "porcelain cat". Its low number of objects was chosen to demonstrate individual localization effects and to inspect specific problems. Of course, the performance is not significant due to the low number of objects, but we evaluate for this purpose the sets with 15 and 100 objects. The test sets were created from completely different data as the training sets to ensure their independence. First, we selected the appropriate number of objects from the COIL-100 database, and then split the images of all these objects 50%:50% in a training and a test image set. The training set was

used for the learning of the system (Sec. 4.2.4), while the test set was used to construct the test scenes.

Concerning the background, we use black backgrounds, white-noise backgrounds similar as in Pinto et al. (2008), and real-world backgrounds. The first class represents a typical indoor laboratory setup as used for example in Antonelli et al. (2014), and is suitable to inspect attention mechanisms without effects from background. The second one (Fig. 4.6a) supports the comparison with other evaluation studies (Pinto et al., 2008), and it is used because it causes a strong noise on the responses of V1 making the recognition harder. Differing from Pinto et al. (2008), we use colored white-noise as our V1 model is excited by color information too. The colored noise is created by adding white-noise independently to each RGB-channel. The third one is used to evaluate the system in a realistic context, hence in a scene in which human subjects have to localize objects similar to the ones from the COIL-100 database. The database consists of man-made objects which are mainly used by humans in their close range or even within-arm length (peripersonal space, Previc (1998)). Hence, we use primarily images of man-made and close range environments as backgrounds. We chose 13 arbitrary backgrounds and ensure that they contain a realistic amount of distracting background clutter like cluttered textures (Fig. 4.6b, “1”), similar colors to the foreground object (Fig. 4.6b, “2”), a wide range of colors (Fig. 4.6b) and different background object sizes (Fig. 4.6b, “3”). Each scene has a resolution of 800x600 pixels as the COIL objects have a realistic size in such a scene.

Test scenes were created by firstly segmenting the objects via masks provided by Parks and Levine (2006), and then placing them randomly in the scene. In opposing to Pinto et al. (2008), we use no additionally transformation of the objects. We choose five objects per scene, one of them serving as the target and four as the distractor. Each object belongs to a different object category to ensure that only one target object is presented. Five objects per scene results in enough distractors, but also avoids crowding effects (Whitney and Levi, 2011). In each scene, each object was localized separately, resulting in five localization tasks per scene. A test set includes 10 scenes per object, thus the sets with 5, 15, or 100 objects contain 50, 150, or 1000 scenes. Therefore, the performance measurements in this section are based on 250, 750, or 5000 tasks. Each task is executed as described in the following. The test scene is presented for 750ms, and the PFC neuron encoding the target object is activated. The test scene is followed by a 250ms blank scene to diminish neuronal activity to zero between the localization tasks. In each task, the model selects an object location via a saccade, or if this fails, it is stopped after the 750ms presentation time.

Thus, the localization process can have four different outcomes: the target is selected, a distractor is selected, the background is selected, or no selection occurs. An object is counted as selected if the saccade target position is either within the object borders or not more than 50 pixel Euclidian distance away from them. This value is roughly half the size of an object.

Performance is measured via localization accuracies ACC and confusion matrices M (Sokolova and Lapalme, 2009). We use the probabilistic variant of a confusion matrix, which shows the probability that a class Y is predicted as class X. In our task, it reflects the probability that the searched target object Y is localized as object X. The matrix is obtained by firstly counting each Y-X pair, and afterwards normalizing each entry with the number of test stimuli of object Y (confusion ratio matrix, Ferri et al. (2003)). In this matrix, the major diagonal represents the probability to select an object l correctly (TPR_l). We define the localization accuracy as the mean over these probabilities:

$$ACC = \frac{1}{\#l} \sum_l TPR_l \tag{4.73}$$

4.4.2. Performance of the full system

The full system can localize quite well the objects with black background (Fig. 4.7a). The accuracies are 92% for the test set with all 100 objects, 96% for 15 objects, and 92% for 5 objects. Localizations errors occur mainly due to a confusion of similar objects, i.e. if a distractor appears similarly to the presented target object (Fig. 4.7d). If the distractor is more salient than the target, its selection is even more likely. For example, the “wooden mushroom” is often selected instead of the presented target “wooden cup”, typically in similar views, which do not show the handhold of the cup and display the mushroom frontally and saliently (top pair in Fig. 4.7d). Thus, the errors depend on the similarity of the two views and so typically occur only for some view points. For the given example, 12% of the presented cup’s were localized as mushroom’s in the 5 object test set, and 14% of the presented mushroom’s as cup’s (Fig. 4.9a in the next section). Such kinds of errors are to a certain degree expectable as also humans make errors under these conditions, i.e. when the target and distractor are similar (hard visual search, see Chap. 2.4). In such cases, humans typically scan several objects until the target is found, thus the here-simulated first saccade is often not directed to the target object.

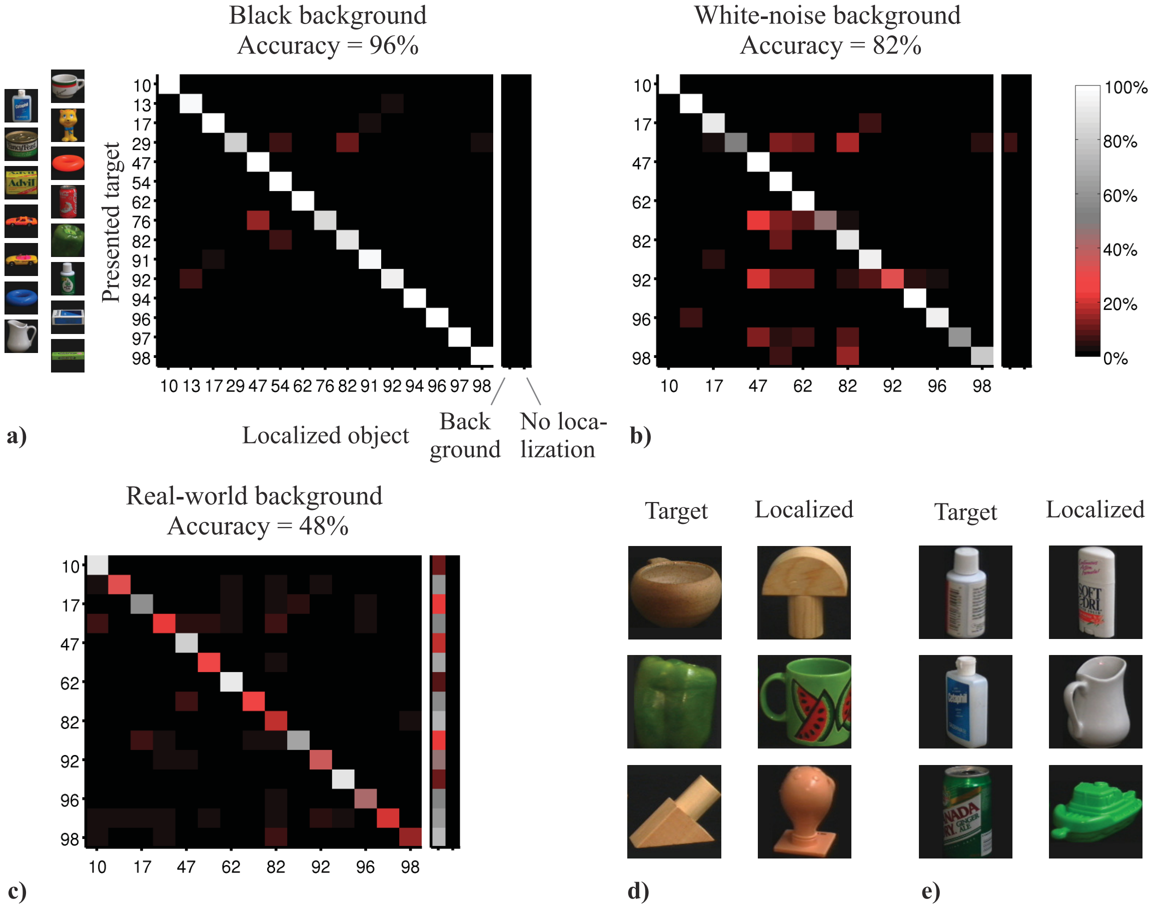


Figure 4.7.: Performance of the full system. **a-c)** The localization accuracies and the confusion matrices (Sokolova and Lapalme, 2009) illustrate the system's abilities to localize objects at the 15 object test set with the black (a), white-noise (b), and real-world backgrounds (c). Low values under 50% are illustrated in red to show mislocalizations. The confusion matrices of the five and 100 objects sets can be found in the Appendix B. For all matrices, the x-axis denotes the localized object or two special cases. The case "Background" indicates that the background was selected by a saccade. The case "No localization" indicates that no location was selected because the system does not execute a saccade. **d, e)** Three examples of mislocalized objects in the sets with black (d) and real-world backgrounds (e). We illustrated the objects without background in precisely the rotation causing the error.

In comparison to other work on the database, our achieved performances of 92% - 96% are higher than in the solid-biologically approach of Einhäuser et al. (2005). They reported 75% accuracy with 50 objects. However, their experiment was slightly different: they use an easier setup with no distractors, but make the training harder by using only 16% of the data set for training, whereas we use 50%. Our accuracies are slightly lower than the less biologically, top-down saliency model of Elazary and Itti (2010). They achieved an accuracy of 97% in a very similar setup as ours, using also black backgrounds and 50% of the data set for training. Their slightly higher performance might result from using a supervised learning approach, whereas our object representations are unsupervised learned.

Localization errors occur if similar objects are represented by the same view cell in HVA. The representation is provided by the external learning procedure (Sec. 4.2.4). However, the procedure does not use the object category information to learn the visual representation (unsupervised learning) as the biological mechanisms of category driven learning in low- and mid-level visual areas are not fully understood. For example it is not known, how the object category information, encoded in PFC (Ashby and Spiering, 2004) and the basal ganglia (Seger, 2008), is projected from these areas to the low- and mid-level visual areas. Instead, the procedure uses temporal continuity information to assembly the representation. Thus, the approach creates the object view representation from consecutively-presented images. However, it has also the preference to learn similar stimuli into one view as it is a modified Hebbian learning rule. Therefore, it can happen that an object view cell is learned from inputs belonging to two different objects if they are similar enough, resulting in a cell representing two objects, which causes the errors.

The generalization to test sets with white-noise and with real-world background results in lower accuracies: 89%, 82%, 71% for sets with 5, 15, 100 objects respectively and white-noise backgrounds (Fig. 4.7b), and 38%, 48%, 42% for sets with real-world backgrounds (Fig. 4.7c). However, this performance drop is not surprising as other studies reported similar results. For example, Lyu (2005) reported on the COIL-100 database a drop from 98.6% to 82.3% by adding Gaussian-noise. They recognized single objects with a support vector machine (SVM, Hearst et al. (1998)). The SVM operates on a stimulus representation consistent of local gray-scaled-based features (jet features, Schmid and Mohr (1997)), which might be comparable to our edge filters in V1. The performance decrease is smaller than in our model, yet they used also a lower amount of noise.

The performance impairments are caused by four reasons. Firstly, the white-noise backgrounds induce noise at the neuronal responses in V1. The main amount of noise is eliminated in HVA because its view-tuned cells react less sensitive to white-noise backgrounds, and because the feature-based suppression reduces neuronal noise (Sec. 4.3.2). However, the remaining amount causes misclassifications in multiple object classes as shown by the confusion matrices for the set with 15 objects (Fig. 4.7a vs. 4.7b).

Secondly, the real-world scenes cause a changed appearance of the target object (Fig. 4.8a-c). The appearance is altered as the object was learned on black background and is now embedded into a real-world scene. This changes for example the contrast of the outer object borders, which changes the neuronal response in V1. The response differs because the V1 filter bank operates on contrast differences, either in form of the color contrasts in the L-M or S-LM channel, or in form of Gabor filters in the orientation channel. All V1 neurons have an excitatory 'on' region and an inhibitory 'off' region. One of these regions is changed by the real-background, resulting in a changed V1 response. Fig. 4.8a illustrates this effect for an object border encoded as a vertical white-black edge. The altered V1 response pattern does no longer match the encoding of an object, learned at black background. This results in a weak excitation in HVA, causing the mislocalization.

Thirdly, the real-world background changes the V1 responses outside the target object. This could also change the HVA excitation as a HVA cell has for implementation purposes always the same receptive field size of 128x128 pixels. This size correspondent to the extent of large objects, and contains black background for small objects. Thus, the weights encode typically a small area outside the object too. The objects were encoded via positive and negative weights. The latter encodes V1 neurons which should not be active for a certain object. This encoding massively helps to distinguish objects (Wiltschut and Hamker, 2009). For example, it helps to discriminate between small and big objects as small objects are encoded via negative weights outside the object borders. In the training phase, the V1 cells encoded by negative weights are inactive. If the target is now presented in the test phase before a real-world background, these V1 cells could become active, which leads incorrectly to a weakening of the target's HVA excitation. This problem can be so severe that the excitation of the target is completely erased. Fig. 4.8b illustrates the case for a small target object encoded via the S-LM channel by plotting feature five in the channel (dark yellow encoded as a weak response of LM without S). This case and the changed border contrast result in a large amount of mislocalizations.

The last reason is the noise from stimuli in the background. Problematic are especially salient stimuli as they cause a strong response in V1 and potentially in HVA. Also problematic are stimuli appearing very similar to the objects in the COIL-100 database as they would trigger a strong response in the HVA cells. Fig. 4.8d shows four examples of mislocalized objects.

To conclude, our model achieves good localization accuracies of around 95% at the test sets with black background. This performance is in the range of other biological models

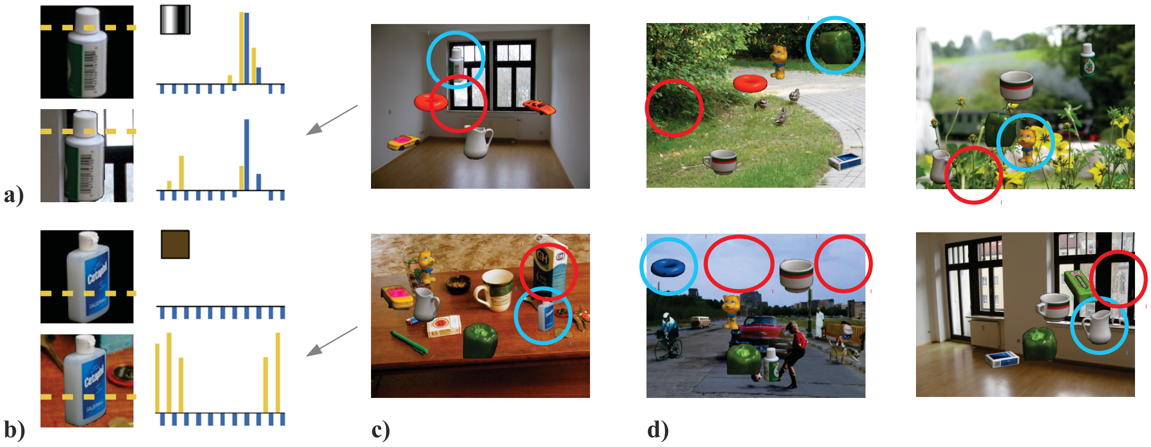


Figure 4.8.: Localization errors caused by real-world backgrounds. **a, b)** One source of localization errors is a changed appearance of the target object. The upper panels show the appearance of the target object during the learning phase (left), and its encoding in a representative V1 feature map (right) along a slice through the object (dotted yellow line on the left). The feature map encodes a vertical white-black edge in (a) and the color dark yellow in (b). Yellow bars denote the neuronal responses in V1, and blue bars the object encoding via the weights V1→HVA layer 4. The lower panels show the same objects in the test phase before a real-world background. A real-world background can change the V1 responses at the object borders (a) or outside the object (b). The V1 responses now mismatch with the weights, so the HVA neurons are not excited, leading to the mislocalization. **c)** Examples of mislocalized objects in the real-world scenes due to the changed target appearance. In each scene, the target object is marked with a blue circle and the incorrectly selected background region or stimulus with a red circle. The examples in a), b) are taken from these scenes. **d)** Another source of localization errors are background stimuli appearing similar to the target object. We give here four examples, which are similarly denoted as in c).

and errors occur under expected conditions. If the model has to generalize to white-noise or real-world backgrounds, the accuracies decrease to around 70% and 40% respectively. An analysis of the localization errors shows that they stem from several reasons, for example from a suboptimal object representation which sometimes cannot distinguish between similar objects, or from the background which changes the appearance of the target object.

4.4.3. Performance impact of the feature-based amplification

We now evaluate the impact of the feature-based amplification mechanism on the behavioral performance by disabling it. Amplification represents the task target in the model by amplifying all neurons encoding it (Sec. 4.3.1). Thus, disabling the mechanism should destroy the task performance, which we verify in the following.

As expected, the model localizes incorrectly the objects (Fig. 4.9a vs. c). It selects the most salient object in a scene instead of the target object, for example the "wooden mushroom" in the scenes of the five objects set. We illustrate the result on this set as all scenes contain the same objects. Thus, the confusion matrix shows nicely the selection of the most salient object. In the other sets, the salient object differs across scenes as the objects in each scene were randomly chosen from 15 or 100 objects, blurring the effect in the confusion matrix.

To clarify the performance impact of the amplification more finely, we decrease the amplification strength linearly from 100% to 0%. On the black backgrounds, the decreased amplification does not impair the performance until 50% of its maximum value, but then diminishes it whereby the strongest drop occurs at values lower as 25% (Fig. 4.9b). This indicates that the amplification must be simply high enough to keep the system correctly operating. The same effect is observed for 5 and 15 objects, except that the drop starts at slightly lower amplification levels. The accuracy rates drop with black backgrounds for all object numbers to 20%. The same is observed for white-noise backgrounds, but not for real-world backgrounds sets in which the accuracies drop to 9% for the 5 object test sets, and to 5% for 15 and 100 objects. The value of 20% represents the chance level to localize an object correctly as each scene contains five objects. However, this chance level calculation assumes that the model selects in each test scene one of the five objects. Thus, the two special cases never occur, i.e. the system never selects the background and it always executes a saccade. These assumptions are valid for the black as well as for white-noise background sets, so the accuracies drop to chance level. Yet, they are invalid for the real-world sets as the background is also often selected, explaining the lower accuracies.

The white-noise and real-world background tests are impaired by a greater amount from decreasing the amplification (Fig. 4.9d, e). A non-linear dropping is observed for all amplification strengths, whereby the greatest decline occurs between 50% - 0%. However, a sharp decrease like in the black backgrounds was not observed. This shows that the white-noise and real-world backgrounds need a higher amount of amplification for optimal accuracy. A closer investigation reveals the reason behind this observation: the

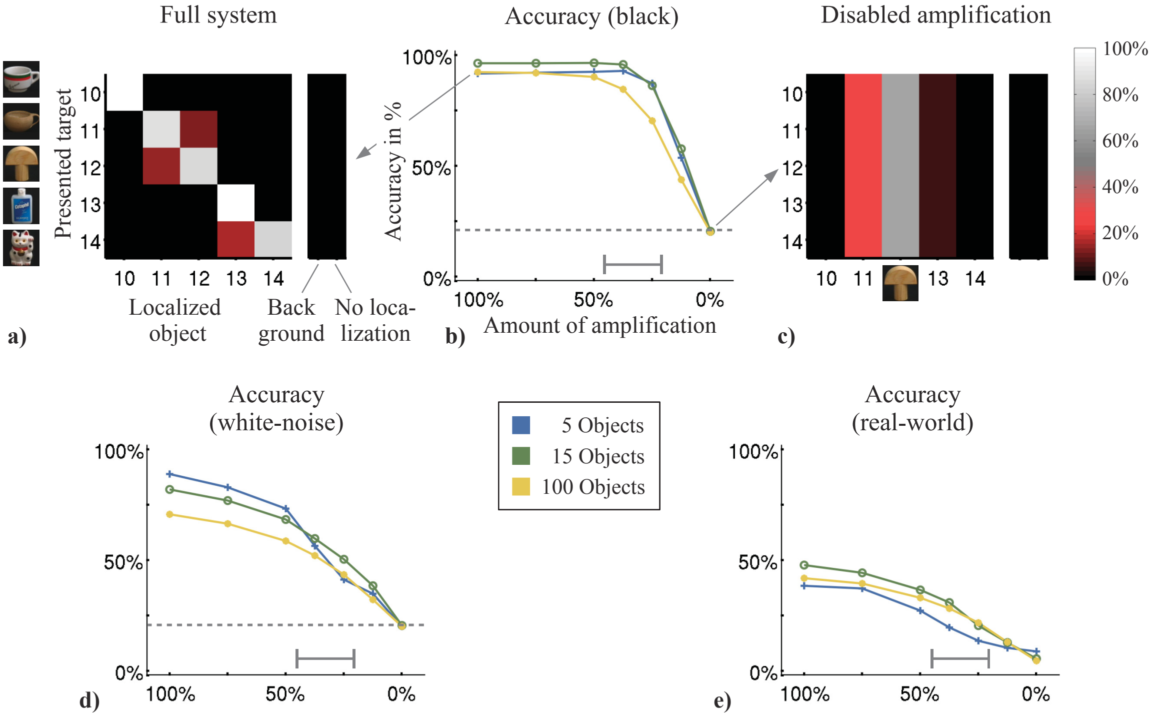


Figure 4.9.: Performance impact of the feature-based amplification. **a-c)** The performance impact on the black background test sets. Two confusion matrices illustrate the localization abilities of the full system (a) and of the system without amplification (c). Both matrices show the performance on the five objects set and are denoted identical as in Fig. 4.7a. b) Decreasing the amplification strength from 100% to 0% impairs the localization accuracies. The two confusion matrices illustrate the extreme cases with 100% and 0% amplification strength. The dotted line in b) and d) denotes chance probability. The bar with flankers indicates the value range used in fitting neurophysiological data sets of attention (46% - 20%). **d-e)** The impairment of the accuracies at white-noise (d) and real-world backgrounds (e). Figures are illustrated identical to b).

amplification must increase for a correct localization the neuronal target signal above the noise signal (Sec. 4.3.1). Unfortunately, the target signal is decreased if the target appearance is changed by the background, and the noise signal is increased if the real-world background is salient and similar to an encoded object (Sec. 4.4.2). Thus, the required amplification level will be higher. The precise level depends on the target and the background, thus it greatly differs among test scenes. When we decrease the amplification, the number of scenes increases gradually in which the amplification level is insufficient for a correct localization, explaining the gradual decline of the accuracy curves. Therefore, the amplification must be higher in white-noise and real-world backgrounds to provide for all scenes a sufficient enhancement of the target signal.

An interesting picture is formed by relating the amount of amplification in the current object localization experiment to recorded neurophysiological data. Chap. 3 fitted a large range of physiological data sets with the mechanistic microcircuit model of attention. This microcircuit model is included in the proposed model in form of HVA. Both models use the same parameter v^{A} for setting the strength of the feature-based amplification. This allows us to related the parameter values fitting the physiological data to the values used in the object localization task. On black backgrounds, the parameter range of the physiological data starts at the minimum value before the accuracy begins to drop (50% in Fig. 4.9b), and ends at the value before the performance dramatically drops (<25%). Thus, the physiological data represents the lowest possible amplification level still ensuring a good performance. Regarding the white-noise and black backgrounds, the physiological data covers a range with accuracies lower as the optimum. These findings indicate that the physiological amplification factors are too low for white-noise and real-world scenes, but are suitable for black background scenes. This conclusion matches nicely the fact that the physiological data was obtained in setups similar to the black background scenes as they use also a black or grayish background. Reciprocally, it predicts that in physiological experiments with white-noise or real-world backgrounds, a much higher amplification level would be observed. Unfortunately, as physiological experiments have not used to the author's knowledge such backgrounds yet, there is no data available to verify this prediction.

To conclude, attention performs the object localization task via this mechanism by amplifying the features of the target object. Hence, the localization accuracy is severely impaired if the mechanism is disabled. A gradual deactivation at the black backgrounds test sets shows almost no impairment until a certain threshold is reached. The threshold matches nicely the value range to fit the neurophysiological data. The white-noise and real-world background sets show already an impairment from the beginning, thus they require a higher amount of amplification for maximal performance. This predicts that a higher amount of amplification might be found in physiological experiments using such backgrounds.

4.4.4. Performance impact of the feature-based suppression

We will now investigate the impact of the feature-based suppression mechanism on behavioral performance by disabling it. Section 4.3.2 explains that the mechanism does not affect localization accuracy, but removes neuronal noise. We will investigate if this is also the case in our large scale test sets, and evaluate the amount of noise removal. Beforehand, we will define our measurement of neuronal noise.

Method to measure neuronal noise

Neuronal noise is defined according to section 4.3.2 as every neuronal activity which is not related to the target object. To measure this property, we record all neurons in the top layer of the model, i.e. in HVA layer 2/3. We divide the recorded neurons in subsets of target and noise related neurons. Target neurons are all neurons representing a feature of the target object at the location of the target. The feature characteristic includes all views associated to the target object, which are represented by the connections PFC→HVA layer 2/3. The location characteristic includes all neurons which receptive field center is located within the borders of the target object. After determining this target set, we assign all other neurons to the noise set. The response of all "noise" cells in HVA layer 2/3 will be shown as a histogram (Fig. 4.10). Inactive neurons will not be shown, thus all neuronal responses lower as certain value, here chosen as 0.15, are excluded from the histogram. The firing rates were recorded at the end of each localization task, hence at the time point of the saccade target selection, or at the end of the stimulus presentation if no target was selected. We need to measure the noise over a complete test set, thus we accumulate the firing rates over all its localization tasks. As the number of recorded cells largely differs between test sets due to different network sizes and number of tasks, we normalize the histogram by dividing its values by the number of neurons and tasks. The resulting bin values denote the percentage of neuronal responses falling within a particular bin in a single localization task in average.

Additionally, we would like to measure the noise via a single value. We define a noise-level $\bar{R}_s$, which is calculated as the mean over the firing rates of all "noise" neurons in test set s. To compare different test sets, we related the values together by defining a noise-factor ν_s (Eq. 4.74). The factor denotes how much more noise the model contains in a particular test set s compared to a reference test set with the noise-level $\bar{R}_0$. As reference set, we chose the most general test set with the lowest noise level, hence the set with 100 objects and black backgrounds.

$$\nu_s \quad = \quad \bar{R}_s \, / \, \bar{R}_0 \tag{4.74}$$

Performance impact

The noise is successfully removed by the feature-based suppression as a low amount of noise can be observed in the full system (Fig. 4.10a). Reducing the suppression to 50%

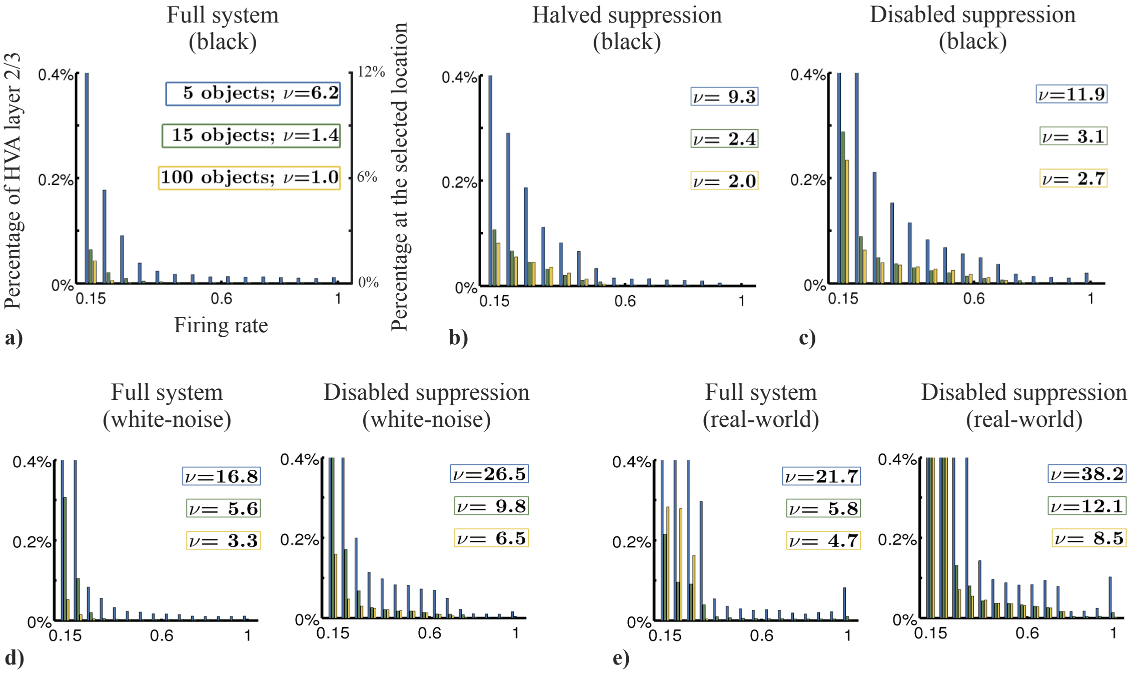


Figure 4.10.: Impact of feature-based suppression on noise removal. **a)** Neuronal noise is defined as every neuronal activity unrelated to the target. The noise is illustrated by a noise-factor ν, and by the firing rate distribution of all neurons in HVA layer 2/3 unrelated to the target. The noise-factor ν describes the increase of noise compared to a reference condition, chosen as the full system with 100 objects and black backgrounds. In the firing rate histogram, neurons without activity are not shown (< 0.15). Neurons are primarily active at the selected location due to the spatial processing. Thus, an additional y-axis on the right side relates the firing rate distribution to the number of neurons in this location, which is around 1/30 of the total number. **b-c)** The neuronal noise increases by: reducing the suppression to 50% (b), or disabling it completely (c). The figures a) – c) show the noise impact on the black background test sets. **d-e)** The neuronal noise in the test sets with white-noise (d) and real-world backgrounds (e). Histograms are similar to a) and c).

increases the amount of noise from $\nu = 6.2$ to $\nu = 9.3$ with 5 objects, showing that some neurons fire at moderate values up to 0.6 (Fig. 4.10b). Disabling the suppression increases the number of firing neurons even more from $\nu = 6.2$ to $\nu = 11.9$ (Fig. 4.10c). Thus, the amount of noise is much higher than in the full system, indicating that indeed the feature-based suppression removes noise by diminishing neuronal activity. Regarding the test sets with a different number of objects, the noise increases by a similar proportion (Fig. 4.10a - c). Regarding the different backgrounds, more neuronal noise is observed in the full system: the factor increases from $\nu = 6.2$ to $\nu = 16.8$ for the white-noise, and to $\nu = 21.7$ for the real-world background test set (five object sets in Fig. 4.10d, e).

Disabling the suppression greatly increases the amount of noise to a much higher level as in the black background tests sets: to $\nu = 26.5$ and $\nu = 38.2$ respectively. This illustrates the effect of the different backgrounds, they induce massively neuronal noise in the system. Concerning the different number of objects, the noise levels increase analogically.

If we compare the absolute noise levels within different numbers of objects, the levels are lower with more objects. The factor ν is in the test sets with 15 objects about 3 - 5 times lower as in the sets with five objects, and with 100 objects about 4 - 7 times lower. This results from a sparser activity of all neurons in HVA layer 2/3 because with a higher number of objects, this area contains much more neurons, but the average number of excited neurons increases only marginally. The sparser activity decreases directly the noise-level $\bar{R}$ as it is based on the average activity in HVA, and so decreases ν as well.

All results show that only a very low number of neurons is active (sparse activity), even without feature-based suppression. This results from the spatial target selection process, which suppresses all non-target locations (Sec. 4.3.2). Therefore, feature-based suppression has the greatest effect in the region of the target location, where it suppresses neurons related to distractors and so removes noise from the neuronal code at the target location. Thus, we add a right-side Y-axis in the histogram (Fig. 4.10) displaying the noise in relation to the number of neurons in the target region only. This axis shows the percentage of noise neurons in relation to the total cell number in the target region.

The localization accuracy should not be affected by disabling the feature-based suppression as it inhibits only weaker neurons, which are negligible for the localization process (Sec. 4.3.2). That is indeed the case as disabling the mechanism changes the localization accuracy by a very low or non-existent amount: for the black background test sets, it alter from 92% to 90% for five objects, from 96% to 96% for 15 objects, and from 92% to 92% for 100 objects. The same picture is formed for the white-noise backgrounds: 89% to 76% for 5 objects, 82% to 80% for 15 objects, and 71% to 70% for 100 objects; and also for the real-world backgrounds: 38% to 36% for five objects, from 48% to 47% for 15 objects and 42% to 40% for 100 objects.

To conclude, we found that the feature-based suppression removes neuronal noise and does not affect the accuracy also in our large scale test sets. If the mechanism is disabled, the amount of noise increases by about $2 - 3\times$, measured by the noise factor ν. Additionally, we evaluated the amount of noise caused by the non-black backgrounds: the noise is about $2 - 3\times$ higher at the white-noise and about $3 - 5\times$ higher at the real-world backgrounds.

4.5. Discussion

In this section, we will discuss existing models, give a survey of the performance boost through visual attention, and discuss improvements as well as restrictions of our model.

4.5.1. Existing models

The proposed attention model is top-down attentional guided, thus we constrain ourselves on discussing top-down models. We will focus on advantages or disadvantages, the role of the attention mechanisms, the evaluation problem, and the reported performance boosts through attention. We would refer the reader to Chap. 2.4 for a general overview of models of visual attention.

Top-down saliency models

The approach adds top-down modulations to the bottom-up saliency models (Borji and Itti, 2013; Frintrop et al., 2010; Filipe and Alexandre, 2013). A typical top-down saliency model is presented by Borji et al. (2009). They extend the pure bottom-up saliency model of Walther and Koch (2006) with learned top-down weights to either amplify relevant or suppress irrelevant feature maps. The top-down attention signals are learned via a global optimization algorithm (comprehensive learning particle swarm optimization, CLPSO, (Liang and Qin, 2006)). Other models operate similar, but use different learning approaches and optimize the top-down signal for different purposes. For example, Navalpakkam and Itti (2006) learn the top-down weights with the goal to maximize the targets saliency relatively to the background, i.e. they maximize the targets signal-to-noise ratio. Many other models belong to the class as a lot of variations exist and they have been applied in numerous real-world tasks. A complete overview is given in Chap. 2.4.2.

These saliency models are classified as top-down, but operate very differently than our model with top-down attention. Our model belongs to the approach of attention as cognitive, holistic control, which simulates the attentional control network of the brain. The top-down saliency models have several disadvantages compared to this approach (Chap. 2.4.3). Firstly, saliency models have the chicken-egg problem that spatial selection and recognition depend on each other. Selection requires information about the object properties and thus a successful recognition, although recognition needs selection to segment the object from the background (Antonelli et al., 2014). Thus, attention and object recognition should be closely intertwined and not separated (Frintrop et al., 2010). In the holistic models, the problem is solved as attention controls recognition and selection in parallel. Recognition relies on the visual areas, which are controlled by top-down attention. Selection relies on the FEF, which operates in parallel to the visual areas. Secondly, the selection in saliency models is based on top-down attention to simple object features, for example the feature ’red’ of a red ball. If a target cannot be distinguished via simple features from background, the selection will fail (Frintrop et al., 2010). Either it selects too few locations resulting in missing recognitions (high true negative rate, Sokolova and Lapalme (2009)), or it selects too many locations which removes the benefit of this stage and might results in many incorrect recognitions (Mitri et al. (2005), high false positive rate). Contrary in the holistic models, attentive signals project to high-level visual areas with neurons encoding whole objects (Logothetis et al., 1995). Thus, a target can be distinguished via these high-level object representations rather than using simple features.

From all top-down saliency models (Sec. 2.4.2), only Borji et al. (2009) and Navalpakkam and Itti (2006) reported the impact of attention. Borji et al. (2009) measured localization accuracies on five real-world test sets. An object was defined as localized correctly if the most salient point was less than 30 pixels away from its border. Localization accuracies were already high without attention at 78% - 83%, and increased with attention to 90% - 94%. This performance gain is rather low compared to our results, however this might be a ceiling effect as the accuracy was already high without attention. Navalpakkam and Itti (2006) evaluated their system by measuring the signal-to-noise ratio (SNR) in the saliency map under the assumption that higher values would result in a better performance. They evaluated four models: a bottom-up one (T0D0), a model in which the target was amplified (T1D0), a model in which the distractor was suppressed (T0D1), and a one with both mechanisms (T1D1). The bottom-up approach performs badly with a mean SNR of -0.77, amplification of the target improves SNR to 2.94, suppression improves SNR to 2.54, and both mechanisms result in the highest SNR of 3.92. These performance gains are qualitatively similar to our results. A quantitative comparison is not possible as the SNR values could not be directly mapped to localization accuracies.

To sum up, top-down saliency models have been widely used, however the approach has some disadvantages compared to attention as holistic control, as employed by our model. The impact of attention is significantly in saliency models, despite it was rarely evaluated.

Models with top-down attention as holistic, cognitive control

Next, we discuss models that utilize attention as the cognitive, holistic control of the visual system. Our model belongs as well to this class. We consider only holistic attention models that can cope with whole objects and real-world scenes: Chikkerur et al. (2010), Walther and Koch (2007), Antonelli et al. (2014), Beuth et al. (2010), Hamker (2005b). This model class simulates the attentional processing of the primate brain, spawning a top-down network from the prefrontal cortex (PFC) to high-level visual areas like IT.

Chikkerur et al. (2010) present a Bayesian inference theory of attention, utilizing both feature-based and spatial attention. This system is similarly structured to our model, it contains an abstract input stage I, and stages corresponding to V4, IT, PFC, and FEF/LIP. Our model contains all of them except we abstract V4 and IT into one stage (HVA). They also use a feedforward processing via I→V4→IT→PFC, and a top-down processing via PFC→IT→V4. The spatial processing is less sophisticated than in our model: they either apply spatial attention to V4 or read out object locations via the signal V4→FEF/LIP, but they have no spatial reentrant processing. To cope with objects and real-world scenes, they process a scene with HMAX (Serre et al., 2005) and set its C2 unit activities as input I. Thus stage V4 will react like C2 to intermediate complex visual features, whereas our HVA encodes view-tuned cells (Logothetis et al., 1995). Both representations have a certain degree of invariance. Such invariances are built-in in HMAX, whereas we learn them based on the temporal information in the retinal image stream (trace learning). As the latter is potentially more advantageously, a recent branch of HMAX also uses trace learning (Isik et al., 2012). Therefore, we see both approaches as appropriate ways to create sophisticated visual representations of the world with a certain degree of invariance. Attention is implemented in this Bayesian inference framework as a probabilistic prior (Congdon, 2007; Edwards et al., 1963). Attention has so the effect to increase or decrease the probabilistic response, which correspondent to the effects of neuronal amplification and suppression in our model. Hence, their approach exploits similar attention mechanism as our model: feature-based and spatial amplification as well as suppression. However, as they always employ the mechanisms of amplification and suppression in parallel, the role of the individual mechanism remained unclear. The model has some physiological plausibility as it can simulate a few physiological attention experiments: multiplicative amplification of tuning function, attentional modulation of contrast response function, and biased competition in visual search. Unfortunately for the physiological comparison, the model simulates probabilities instead of neuronal responses. The authors assume that both are equal, but have to admit in the same paper that it is open question how the brain might represent such probabilities. Thus, we think that this assumption should be taken with care. To sum up, their model has a weaker physiological foundation than ours and lacks spatial reentrant processing. The attention mechanisms and the object representations seem similarly powerful as in our model.

They evaluate their model on a real-world data set, containing cars and pedestrians. The model has either the task to localize an arbitrary car or pedestrian in the image. Thus, it is a two-class problem as the system needs only to distinguish between cars and pedestrians. The model achieves an accuracy of 82% to localize cars and of 81% to localize pedestrians correctly. The authors count an object as correctly localized if the saliency map at the object center exhibits an activity over a certain threshold. This is an easier criterion than in our evaluation relaying on the neuronal dynamics in the FEF as these require that the maximum activity is at the object location. This difference might explain their relatively high localization rates. Interestingly, they show also the impact of attention on the localization accuracy. Without attention, the accuracies were 43% and 39% for cars and pedestrians, with only feature-based attention 68% and 75%, and with only spatial attention 81% and 79%. This shows that attention improves recognition and localization up to a factor of 1.9x (from 43% to 81% accuracy), and that the advancement depends highly on the data set and type of attention. They investigate also the reason for the highest factor and stated that spatial attention helps massively, because the data sets consist mainly of streets scenes which have a strong spatial bias regarding the location of cars and pedestrians.

Walther and Koch (2007) present an object recognition and localization system with top-down attention on the basis of HMAX (Riesenhuber and Poggio, 1999). HMAX is classically a pure feedforward driven recognition architecture consisting of the layers S1, C1, S2, C2, and VTU. The first stage S1 extracts simple features like edges from the image, stage C1 pools them, stage S2 extracts combinations of C1 features like combined edges, C2 pools them again, and VTU (view-tuned units, Logothetis et al. (1995)) extracts views of an object from the C2 response pattern. Walther and Koch (2007) extend this HMAX-hierarchy with layers encoding objects, categories, and tasks. Attention is implemented via top-down feedback connections projecting the hierarchy backwards until stage S2, i.e. via the chain task→category→object→C2→S2. The feedback connections are reciprocal build to the feedforward connections. The attention signal modulates multiplicatively the neuronal activity, thus values greater than 1 result in amplification, and smaller than 1 in suppression. Thus, the model also employs amplification and suppression as one operation, making the individual role of each one rather vague. Attention controls the system for the current task by amplifying neurons encoding task-relevant entities. The amplification starts in the highest layer by amplifying task-relevant categories, and then continues by amplifying all objects belonging to this category, then all visual representations belonging to this object, and so on. Suppression on the other hand inhibits all neurons encoding task-irrelevant entities. In summary, the model combines nicely HMAX with a top-down control via feature-based attention. On the other hand, the model misses a grounding on physiological data and lacks a spatial attentional processing, thus it has some serious disadvantages in relation to our work.

Walther and Koch (2007) evaluate their system by detecting frontal views of human faces in a custom image data base (faces vs. background, i.e. a two-class problem) and by measuring the number of fixations until a frontal view was found. The fixations were calculated via a saliency map, which was extracted from the S2 features. They evaluate a pure bottom-up approach (1), a top-down attention approach with stage S2 learned on arbitrary patches (2), and a top-down one with S2 learned on patches at or around faces (3). Approach 1 needs in 9% of the cases only one fixation, approach 2 in 21%, and approach 3 in 38%. Thus, the top-down attention improves the performance by a factor of 2x - 4x. However, this evaluation has a major drawback: the performance of the basic approach is with 9% rather bad, thus any change in the system will likely improve the result. Thus it is not clear if the improvement belongs to top-down attention or to other factors, like the differently learned S2 stage. Therefore, attention seems to improve the performance in their system, but we cannot validate to which extent.

Antonelli et al. (2014), Beuth et al. (2010), Hamker (2005b) belong to research group of Prof. Hamker. All three models have a comparable structure, similar to our model: lower and higher visual areas of the ventral stream, the prefrontal cortex, and the frontal eye field. Hamker (2005b) simulates an input stage (comparable to V1), a stage 1 (V4), and a stage 2 (IT). Antonelli et al. (2014), Beuth et al. (2010), and our model condense stage 1 and 2 into an abstract “higher visual area” (HVA). The reason is that we assume that the attentional processing in V4 and IT is very similar as several studies have been reported similar phenomena (Chap. 3). In Hamker (2005b), V4 and IT encode an object as a set of low-level features. Antonelli et al. (2014), Beuth et al. (2010), and our model represent an object more sophisticatedly by view-tuned cells (Logothetis et al., 1995), and employ as novelty a trace learning procedure to create such cells. All models contain a top-down feature-based attention signal that originates from PFC and is projected backward to high-level visual areas in the model: V4 and IT in Hamker (2005b), and HVA in Beuth et al.

(2010), Antonelli et al. (2014). Spatial attention is also present in all models via a reentrant processing between these high-level visual areas and the FEF. Attention is implemented as multiplicatively amplification, and is combined with feature-based and spatial suppression mechanisms. Therefore, all models contain all four attention mechanisms. Biological plausibility has been solidly shown for Hamker (2005b) in a separate publication: Hamker (2005a). It demonstrates that the model accounts for neuronal responses of area IT in a biased competition/guided visual search setup, and it predicts the responses of FEF cells in guided visual search. Furthermore, it considers anatomical and physiological constraints of V4 and IT, and it models detailed the different cell types of the FEF (Schall, 1991): visuomovement, movement, and fixation cells. Due to this high biological plausibility, our model is strongly based on Hamker (2005b). We advance it in multiple ways, thus we discuss these improvements in a separate section (Sec. 4.5.3). To summarize, all three models simulate in detail the attentional control network within the ventral stream and the FEF. Antonelli et al. (2014) and Beuth et al. (2010) have as advantage the trace learning, while Hamker (2005b) has a very good biological foundation.

Object localization accuracies are reported in all publications. Hamker (2005b) reported 81% accuracy to fixate the target object correctly in the first four fixations and 50% to fixate it immediately correctly. They used 16 different objects in a data set of three real-world scenes. Beuth et al. (2010) reported a 100% ability to discriminate correctly between 10 different objects in a virtual reality, but this evaluation was done on scenes containing only a single object. Antonelli et al. (2014) reported 100% accuracy to localize 3 objects correctly in 27 different indoor laboratory scenes. In summary, the reported accuracies are really high, but also the tasks were relatively easy due to low number of objects and test scenes. The high accuracies result certainly also from the used amplification and suppression mechanisms, but none of these studies evaluated the impact of these mechanisms.

To conclude, Hamker (2005b) is the most advanced model in terms of biological plausibility (except our model), but represents an object only as set of low-level features. The other models learn more superior object representations, either via HMAX (Chikkerur et al., 2010; Walther and Koch, 2007) or trace learning (Antonelli et al., 2014; Beuth et al., 2010). However, they have a weak biological foundation, and the first two lack additionally a spatial reentrant processing. We based our model on Hamker (2005b) due to its biological plausibility, but further developed it by incorporating the microcircuit and the trace learning from the other models. Thus, our model may constitute currently the most sophisticated one.

4.5.2. Survey of reported performance boosts due to attention

We will now examine the reported performance boosts in our and other models to quantify the benefits of attention for object localization. In our model, we can distinguish between the effects of feature-based amplification and suppression. We found that (Sec. 4.4): 1) disabling feature-based amplification removes the encoding of target features, decreasing the localization accuracy massively at all backgrounds: from 92% to 20% at black, from 71% to 20% at white-noise, and from 42% to 5% at real-world backgrounds. All here reported values belong to the test set with 100 objects. 2) Disabling feature-based suppression results in a higher amount of neuronal noise, leading to an increase of the noise-factor ν of about $2\times$ at all backgrounds: from 1.0 to 2.7 at black, from 3.3 to 6.5 at white-noise, and from 4.7 to 8.5 at real-world backgrounds. The factor ν compares the mean activity of all neurons related to noise with a reference condition (Sec. 4.4.4). Regarding the localization accuracy, the suppression mechanism does not affect it significantly.

Other holistic models reported only the combined accuracy boosts of amplification and suppression (Sec. 4.5.1). Chikkerur et al. (2010) reported an increase of the localization accuracy from 43% to 68% with feature-based attention, to 81% with spatial attention, and to 82% with both forms. Walther and Koch (2007) reported an accuracy increase from 9% to 38%. Previous holistic models of the research group reported only the accuracies with attention: Hamker (2005b) reported 50% and Antonelli et al. (2014) 100%, but the latter on a data set with only three objects.

In the class of top-down saliency models (Sec. 4.5.1), Borji et al. (2009) reported an accuracy increase from 78%-83% to 90%-94%. The accuracy boost stems in our model from the amplification mechanism. As the other models include a similar mechanism too, we speculatively relate the boost to the amplification.

Quantitatively, the reported numbers differ, but they all show a huge performance increase of about 20% - 70%. Only Borji et al. (2009) shows a smaller amount of about 10%, which might be a saturation effect as the accuracy without attention was already high. Alternatively, it might indicate that attention is less crucial for top-down saliency models than for holistic models. The baseline performance without attention differs strongly between the studies, which might be caused by a more salient target in one study than in another. When attention is disabled in our model, the most salient object is selected instead of the target. We assume a similar behavior in the other models. If the target is more salient than all distractors in a particular task, the localization will be correctly even without attention. Hence, the performance without attention in a particular study depends strongly on the target's saliency in relation to all distractors. Besides this reason, the reported variations are to a certain degree expected as all studies use different data sets and evaluate different models. Nevertheless, this survey gives an estimate about possible performance boosts through attention.

4.5.3. Improvements in comparison to previous works

The proposed model stems primary from the model of Hamker (2005b), but was improved as outlined in the following. Specific aspects of this model were enhanced in the recent years by the author and published separately (Antonelli et al., 2014; Beuth et al., 2010; Zirnsak et al., 2011a), (Chap. 3). We now reintegrate them into the proposed model, leading to novel, general system-level model of attention.

As a first improvement, our model uses learned object representations. The previous model of Hamker (2005b) uses hand-crafted low-level feature extractors (saliency model, Itti et al. (1998)) and does not recombine them to object representations. Thus, the system could only distinguish objects based on low-level features like color, illumination, or global orientation of an object. The feature "global orientation" encodes the global average orientation of the object, but not its shape. Therefore, the system could not distinguish similar objects and so was limited to recognize very different ones. In contrast in this work, the weights from V1 to HVA were learned by the trace learning approach to encode object views (Sec. 4.2.4). The learning approach is theoretically usable for all kind of objects (Antonelli et al., 2014). These learned object descriptors were also used in Antonelli et al. (2014) and Beuth et al. (2010), the previous works of the author. Antonelli et al. (2014) demonstrate the learning of three objects in laboratory scenes containing a dark-grayish background, and Beuth et al. (2010) of ten objects before black backgrounds. The proposed model advances these studies by learning a very large number of 100 objects and by using real-world backgrounds. Our satisfying results with 100 objects (Sec. 4.4.2) indicate that the system can work with almost any number of objects.

Compared to Hamker (2005b), neurons in HVA have now cell properties similar as in the high-level visual cortices. Logothetis et al. (1995) found that cells in area IT react to specific views of an object. Such cells emerge directly from the learning process. The HVA encoding provides also now a more object specific search template. Each HVA cells prefers an object from a particular view points, thus the receptive field of each cell can be seen as a template for a particular view. Views are associated to an object via the PFC→HVA connection. Via this connectivity, a search for a certain object will result in activating multiple HVA cells, which together provide a search template as the superimposition of the receptive fields of these cells.

The receptive field of a HVA cells determines which V1 low-level features are bound together to form an object representations. This is known as binding problem (Hamker, 2005b; Rothenstein et al., 2008). The learning algorithm forms the receptive fields based on the temporal continuity principle, and so solves the binding problem. The preceding model of Hamker (2005b) proposes for this purpose spatial attention, which amplifies spatially adjacent neurons, and so bind their low-level features together. In our model, spatial attention amplifies similarly adjacent neurons in HVA, however this binding plays a different role. It binds neurons at adjacent locations together which help to focus neuronal activity on the center of an object.

The trace learning is supported by Anti-Hebbian learning to create a pattern of suppressive connections between HVA cells, known as lateral inhibition (Ratliff et al., 1967). We investigate in the following if this pattern is consistent with the neurophysiologically-based connectivity in the microcircuit model of attention (Chap. 3), and whether it is suitable for attentional processing. The Anti-Hebbian learning (Antonelli et al., 2014; Wiltschut and Hamker, 2009) implements a competition between encoded features. The competition is strengthen if features are activated at the same time. This learning leads to a strong competition, i.e. strong connections, between similar features and to a weak competition between dissimilar ones. Yet, the microcircuit model predicts the opposite connectivity pattern: non-existing connections between similar features and strong ones between dissimilar ones. The reason is that the modeled neurophysiological data shows no competition between similar features and a strong competition between dissimilar ones (Lee and Maunsell, 2010a; MacEvoy et al., 2009). Therefore, Anti-Hebbian learning produces a connectivity pattern contradictory to the one predicted by neurophysiological findings. Furthermore, we tested the Anti-Hebbian connectivity pattern in our system-level model of attention, and found that it results in a sub-optimal working model. The strong suppression between similar features occurs often between views belonging to the same object, resulting in a strongly reduced HVA response towards this object and impairing the accuracy. The weak suppression between dissimilar views does not sufficiently suppress neuronal activity resulting from distractors and background. Due to these reasons, we model the suppressive connections according to the microcircuit: they are zero between views belonging to the same object, and strong between all other views. To sum up, the Anti-Hebbian learning produces a suppressive connectivity pattern that is neither supported by neurophysiological data sets of attention nor suitable for attentional processing.

The model of Hamker (2005b) had already a very good neuroanatomical and physiological background (Hamker, 2005a), however we improved it with recent discoveries and model approaches. The two HVA layers are now simulated by the microcircuit model of attention (Chap. 3), a model which can account for a large data set of physiological attention experiments: spatial and feature-based biased competition, modulation of the contrast response function, modulation of the neuronal tuning curve, and modulation of surround suppression. The FEF was also improved, regarding its neuroanatomical foundation, by integrating the advanced FEF model of Zirnsak et al. (2011a). According to Pouget et al. (2009), Hamker (2005b) did not take into account that the FEF to V4 projections may originate anatomically from visuomovement cells (FEFvm), and not from the FEF movement cells (FEFm) as implemented in the model. A projection from the FEFvm cells is indicated by the anatomical observation that the FEF to V4 connections originate from supragranular layers in the FEF (i.e. layer 1-3), which contain the FEFvm cells (Zirnsak et al., 2011a). Functionally, this implies that the projected information may contain visual as well as movement related components. The FEF was improved in Zirnsak et al. (2011a) by introducing a new FEFvm cell population, which projects back to V4. The new FEFvm population simulates cells with a varying continuum of visual and movement influences, thus now both components are projected backwards. Additionally, it was added in the model that the visual and visuomovement cells types of the FEF are simulated in separate populations. The physiological literature distinguishes between both cells types, but this was not taken into account in Hamker (2005a), thus it was added during the updating of the FEF model. By incorporating these two models into our model, we strongly improve its biological underpinning.

To conclude, we improve the model of Hamker (2005b) in several ways. As first, we learn view-tuned cells with the trace learning algorithm. The view-tuned cells represent an object more sophisticatedly than the set of low-level features used previously, and HVA contains now cells as in high-level visual areas like IT. Moreover, we show that the view cells of an object provide a sophisticated search template, and show that Anti-Hebbian learning produces a connectivity pattern unsuitable for attentional processing. Finally, we increase the already good biological foundation even more by incorporating the microcircuit model and the improved FEF model into our work.

4.5.4. Limits of the proposed model

The model has a few restrictions as outlined in the following. At first, the objects can be recognized under different rotations, but not under other transformations. This is due to our focus on learning rotation invariance. However, Antonelli et al. (2014) show that the approach can learn to recognize objects under many different transformations, due to its view-based representation (Beuth et al., 2010). To recapitulate, a view represents a specific visual appearance of an object and when a transformation changes the visual appearance, a different view cell will respond. The learning algorithm creates view cells for all transformations included in the learning set. As this recognition approach is completely independent of the kind of the transformation, it is capable to learn arbitrary ones. Here, we showed the approach coping with rotation transformation by including only them in the learning set, while Antonelli et al. (2014) showed it for coping with scaling and with disparity transformations. Other ones could be easily added.

At second, the model assumes that the target object is always present in the scene, as it could not detect the presence of an object. The preceding model of Hamker (2005b) contains for this purpose match detection neurons. They compare the neuronal activity in the highest visual area (stage 2) with the search template of the target object, and report a recognition only for a successful match. However, the activity in this stage 2 represents low-level features, thus the model verifies each low-level feature separately. This mechanism could not be implemented in our model as our HVA neurons encode object views and not low-level features. A solution might be to evaluate the excitation of HVA target neurons, and report only a match if the excitation exceeds a certain threshold. The HVA cells react very specific, thus the excitation will be close to zero if the target is absence. However, the biological realism of this solution is unclear as the cortical mechanisms of a match operation are unknown, thus we leave this point open for the future research.

The next limitation concerns the execution speed of the model. The model is implemented via ordinary differential equations (ODEs, Atkinson (1989)) with the explicit Euler approach (Atkinson, 1989) and a step size of 1 ms. This approach is very simple, but not very fast as it requires an evaluation of the ODEs every simulated millisecond. A single image is processed either until an object was selected by the model, or until the next image is presented. The former criterion takes at least 150 ms simulation time and the latter condition occurs after 750 ms, thus the processing of an image takes between 150 and 750 ms, corresponding to 150 - 750 evaluations of the ODEs. Additionally, 250 evaluations are necessary to presenting 250 ms long a black image afterwards. In contrast, computer science approaches (HMAX, Riesenhuber and Poggio (1999); convolution networks, LeCun et al. (1998)) are typically pure feedforward and hence need only a single evaluation. The proposed system was implemented in MATLAB with C++ and with multithreading via OpenMP (Chapman et al., 2011). In this implementation, a single evaluation needs at average 90ms execution time, thus the processing of a whole image takes between 36s and 90s. This was measured on an Intel i7-3930K processor with 6x-multithreading, an image resolution of 800x600 pixels, and the network configuration for the 100 object set.

The object representations were learned offline, i.e. a separate routine was used to learn the weights, and these learned connectivity matrices were loaded into the attention model. This is clearly contrary to the processing in the primate brain. However, we choose this approach because it allows a much faster processing due to weight sharing. Weight sharing means to learn the connectivity for a small patch holding only one object and then share these weights among all location to allow the processing of whole scenes. Weight sharing would not be possible with learning inside the attention model as it processes whole scenes. All HVA locations would receive different visual inputs, and so would learn different weights. Therefore, the weights could not be shared. Without weight-sharing, the memory usage would increase by the spatial resolution of HVA layer 4, thus here by a factor of 4524. This would result in up to 24 GB of ram in the network configuration for the 100 objects sets, making it no longer possible to run the model on normal desktop PCs. As this memory has to be accessed in each evaluation step, the execution speed would be slower by a certain factor. The factor depends strongly on the size of the weight matrix, on cache effects, and on the cpu architecture. We conducted some preliminary tests and measured for our model a slow down by a factor of 1.5x - 3x. The second benefit from offline-learning on single patches is that the learned features originate from one object. If multiple objects would be in the receptive field of a HVA cell, the learning would incorrectly learn features from multiple objects into a single representation.

Next, the model's object representations are not optimal as the localization errors primarily stem from them and not from the attentive dynamics (Sec. 4.4.2). The representations emerge from the trace learning procedure, which was chosen due its high biological plausibility and its simplicity. However, these properties limit also its performance. Firstly, the procedure does not exploit the category information (unsupervised learning approach) as the biological mechanisms of category driven learning in low- and mid-level visual areas are not fully understood. For example it is not known, how the object category information, encoded in PFC (Ashby and Spiering, 2004) and the basal ganglia (Seger, 2008), is projected from these areas to low- and mid-level visual areas. Thus, the resulting object representations cannot distinguish between very similar objects. This is the main source of the errors on black backgrounds (Sec. 4.4.2). Secondly, we choose the learning approach for its simplicity as we would like to focus our efforts on investigating attentive dynamics and not on the learning. That decision includes to choose an approach that performed well in previous models of our group (Antonelli et al., 2014; Beuth et al., 2010), and includes to learn only on black backgrounds. Yet, the resulting representations does not match with the stimuli before white-noise and real-world backgrounds, and thus are the main source for errors on these backgrounds (Sec. 4.4.2). We currently plan a follow up study to learn on all backgrounds. Therefore, the object representations are relatively simple as we focus on simplicity and biological plausibility of the learning.

As last limitation, the model contains only one area of the visual cortex that is attentional modulated. The visual cortex on the other hand shows attentional modulation in many areas like V1, V2, V4, MT (Saenz et al., 2002), or IT (Desimone, 1998). We model only one area to keep the model simple as other areas operate in a similar way (Chap. 3). Also it seems challenging to use multiple areas as this would require to learn them. The learning of multiple areas seems a currently unsolved problem in computational-neuroscience community because models using a form of temporal Hebbian-Learning typical use only one stage (Földiák, 1991; Spratling, 2005; Teichmann et al., 2012). A recent version of the networks from Spratling can learn two stages (Spratling, 2012), but it is unclear if this model version can operate with temporal Hebbian-Learning. Others like HMAX uses multiple stages but learn only the last stage (Riesenhuber and Poggio, 1999), or learn each stage only via a very simple imprint mechanism (Serre et al., 2007b). In the field of machine learning, deep neuronal networks (LeCun et al., 1998) learn within multiple stages (Krizhevsky et al., 2013; LeCun et al., 2015; Sermanet et al., 2014; Zeiler, 2012), but these networks lack biological plausibility. These findings implicate that the learning over multiple stages is only suboptimal solved in the computational-neuroscience community, and thus we decided to not address this problem here.

4.6. Conclusion

This chapter illustrated how the primate brain localizes a target object in a scene, and shed light on the role of attention in this task. In the brain, attentional processing spawns a top-down control network. This and other recent findings led us to propose a novel, more general view of attention (Sec. 4.2.3): We see it as a cognitive and holistic control process, tuning the visual system for the task at hand. The control is termed as holistic as it modulates neuronal activity in parallel across the whole visual cortex.

We developed here a novel system-level model of attention to improve existing models (Sec. 4.5.1) and to simulate the attentional processing with our previous-elaborated neuronal mechanisms of attention (Chap. 3). Only a few existing neuro-computational (or holistic) models of attention have been applied to real-world object localization tasks. From them, Hamker (2005b) is the most advanced model in terms of biological plausibility, thus we based our model on it. In relation to it, we conducted the following improvements (Sec. 4.5.3): We incorporated our neuronal mechanisms of attention to enhance the physiological plausibility of the attentive dynamics, employed a recent model of the FEF (Zirnsak et al., 2011a) to factor in the anatomical finding that the FEF visuomovement cells project to V4, and learned object descriptors via a trace learning procedure to deal with a high number of objects. We propose our model as an alternative to the classical computer vision approach using attention, i.e. saliency models, as it avoids their drawbacks due to its holistic nature (Sec. 4.5.1).

Our model is demonstrated on large and realistic object localization tasks (Sec. 4.4) to show that neuro-computational models are applicable to realistic computer vision problems. We evaluated the model's accuracy to localize objects on several large test sets using 5, 15, or 100 objects, 1000 scenes, and black, white-noise, or real-world backgrounds. Our model can cope with such high number of objects as we introduce learned object representations, which then are modulated by attention. This approach shows how attention can guide vision with a high number of object categories. The model was learned at the black backgrounds and performs well on them with 92% localization accuracy. Errors occur predominantly when the target and distractor are similar, a case in which also humans execute incorrect saccades, thus the errors are expectable to a certain degree. When the model has to generalize to white-noise or real-world backgrounds, the accuracy diminishes to 71% and 42% respectively. All here reported values were measured on the 100 object sets.

Next, we explain the attentional processing in object localization and analyze the neuronal mechanisms underlying attention for a better understanding. Our model explains object localization as follows (Sec. 4.2): The visual scene is processed by a hierarchy of visual areas, which encodes increasingly complex features, for example objects views in higher visual areas. When a specific object should be searched, feature-based attention is deployed to this target object, simulated in the model by activating a cell encoding the target object in PFC. This activity pattern results in a top-down attention signal to the higher visual areas, which amplifies there all cells encoding a view of the target object, while other cells are suppressed. Thus, cells sensitive for the target object (or parts of it) are higher activated, and hence the neuronal activity in the visual areas is higher at these locations. Spatial processing over the FEF relies then on this biased activity pattern to focus activity on the best of these potential target locations (spatial attention), and in such a way determines the location of the target object.

Regarding the mechanisms, we explained and evaluated the roles of our known neuronal attention mechanisms in object localization (Sec. 4.3 and 4.4 respectively). This firstly clarifies their individual roles in object localization and thus within the holistic control process; and it investigate if all of the mechanisms, which are known from neurophysiology, are relevant for object localization. Linked to this, it secondly clarifies the relevance of a particular mechanism for the localization performance, i.e. how much it contributes to the performance. We focused here on the mechanisms of feature-based amplification and suppression, and found out: 1) The feature-based amplification mechanism performs the enhancement of the target neurons, and thus represents the target in the visual system for the top-down control. Other mechanisms like feature-based suppression or the spatial processing can then rely on this biased activity (Sec. 4.3.1, 4.4.3). The strength of the required amplification signal depends on the background, i.e. it has to be stronger for white-noise and real-world backgrounds. Disabling this mechanism decreases massively the localization accuracy, e.g. on black backgrounds from 92% to 20% (chance level), and the system selects the most salient object or background region. Thus, we conclude that the mechanism is essential for object localization. 2) The feature-based suppression mechanism removes neuronal noise (Sec. 4.3.2, 4.4.4). Noise is defined as all neuronal activities not related to the target, and emerges for example from the background or distracting objects. It is measured by a noise-factor ν, which is 2.7 times higher on black backgrounds if the suppression is deactivated. Interestingly, disabling the mechanism does not impair the localization accuracy (Sec. 4.3.2), thus we found out that the mechanisms is irrelevant for the behavioral performance in object localization. As we will see in the next chapter that this mechanism is very relevant for the behavioral performance in a recognition task, we think its behavioral role may lie more in such tasks, whereas it has a primary neuronal purpose in localization tasks as its removes neuronal noise and thus cleans up the neuronal code.

Finally, we gave a survey of the performance boosts through attention in our and other models as the benefits of attention have been rarely quantified (Sec. 4.5.2). In our holistic model, we observed a large boost of the localization accuracy from 20% to 92% (a boost of 72%), because attention is so crucial for our model that it does not operate correctly without it. Existing studies about holistic models (Chikkerur et al., 2010; Walther and Koch, 2007) have similarly reported a large increase of the localization accuracies of about 20% - 40%. Contrary, the only existing study about a top-down saliency model (Borji et al., 2009) has reported a small boost of about 10%, which might be either a saturation effect or indeed indicating that attention is less crucial for top-down saliency models than for holistic models. Besides, the amount depends also on the data set, e.g. more salient targets cause a smaller boost. Therefore, the reported boosts showed that attention enhances the performance in all cases, but also that the amount depends strongly on the model class and data set.

Future work might improve the moderate accuracies at the white-noise and real-world backgrounds. One reason for these accuracies is the generalization from a black to a different background, altering the appearance of the target object (Sec. 4.5.4, 4.4.2). The object representations were learned on a pure black background, and thus no longer match the changed appearance on a white-noise or real-world background, causing the error. Thus, we are currently conducting another study in which we include white-noise and real-world backgrounds in the trainings set. The learning algorithm might then be able to learn object representations suitable for all backgrounds.

5. The relation of object substitution masking (OSM) and attentive dynamics: A neuro-computational modeling study

Abstract Although object substitution masking has been discussed being affected by attention, OSM is classically not considered emerging from attentive dynamics and there is little overlap between both fields of research. However, we demonstrate here by means of a neuro-computational modeling study that OSM can be fully explained by attentive dynamics. The model was developed to explain attention paradigms (Chap. 3 and 4) and includes the ventral stream as well as the frontal eye field (FEF). After showing that the model reproduces typical OSM data sets, we explain OSM by three neuronal attention mechanisms: feature-based suppression, amplification, and a delayed onset of spatial attention. The first mechanism, feature-based suppression, explains in the attention paradigm of biased competition the suppression among stimuli via inhibitory connections. In OSM, this mechanism realizes a suppression between target and mask, and thus accounts for the declining behavioral performance with longer mask durations. The second mechanism resembles the amplifying effect of attention on the neuronal responses. In OSM, this amplification mechanism maintains the target response after stimulus offset. The last mechanism accounts for the observation that OSM requires a high number of distractors (set size effect) like in visual search paradigms. Our model explains this by a delayed onset of spatial attention that diminishes the influence of the underlying spatial amplification mechanism. Spatial attention emerges intrinsically from the recurrent processing between ventral stream and FEF. Contrary to existing theories of OSM, our attention model is grounded on a large set of physiological and neuroanatomical data. Hence, it provides the opportunity to profoundly address unclear aspects of OSM, which we conduct here regarding the necessity of reentrant processing and the relationship between mask duration and set size. To conclude, OSM and attention paradigms rely on the same underlying mechanisms, indicating that there is no need for a separate OSM theory.

5.1. Introduction

Visual masking characterizes paradigms to decrease the visibility of a target with a mask stimulus (Breitmeyer and Ogmen, 2006). It is an interesting psychophysical phenomenon on its own, but is often used to explore the temporal dynamics of visual processing.

Here we will investigate object substitution masking *(OSM)*, which was firstly described by Di Lollo et al. (2000). In OSM, a target, a four-dot mask, and several distractors are briefly displayed together. Afterwards, the mask remains alone on the display, denoted as *trailing mask*. The task is to recognize the target surrounded by the four dots, thus the four dot pattern serves as mask and as target indicator simultaneously. The target recognition is severely impaired when the mask remains on the screen *(trailing mask condition)*, but not if the mask is set off at the same time as the target *(common offset condition)*. For the subjects, the target seems to appear as the mask, which leads to the name of the paradigm as the target seems to be substituted by the mask (Di Lollo et al., 2000).

OSM is broadly reviewed in Goodhew et al. (2013), so we will focus on only two specific aspects of OSM: the relation to other masking paradigms (Breitmeyer and Ogmen, 2006) and the major approaches to model it. OSM is mostly related to metacontrast masking (Shelley-Tremblay and Mack, 1999; Boyer and Ro, 2007) as both have two common aspects: The target appears before the mask, and the mask is displayed spatially adjacent to the target (Di Lollo et al., 2000). However, OSM also differs from it (Di Lollo et al., 2000; Enns, 2004; Goodhew et al., 2013) in three ways: Firstly, the four dot mask is much sparser (Goodhew et al., 2013), and its shape does not follow closely the target shape. Both properties rule out contour interactions according to Di Lollo et al. (2000). Secondly, the mask is typically set on after the target in metacontrast masking (SOA law, Breitmeyer (1984)), whereas in OSM, the stimuli are set on simultaneously (common onset). In such common onset conditions, OSM shows a masking effect, while no effect is observable in metacontrast masking (Di Lollo et al., 2000). Finally, the four-dot pattern in OSM is used as the target indicator and as the mask, whereas metacontrast masking uses typically different patterns.

As reviewed by Goodhew et al. (2013), OSM can be explained either by feedforward / reentrant theories (Bridgeman, 2007; Di Lollo et al., 2000; Francis and Hermens, 2002; Põder, 2013; Smith and Ratcliff, 2009), or by the object updating hypothesis (Lleras and Moore, 2003; Moore and Lleras, 2005; Pilling and Gellatly, 2010). From the first branch of theories, we will focus on the computational model of object substitution (CMOS) by the original discoverer of OSM, Di Lollo et al. (2000). This reentrant model is relatively popular and one of the few computational models of OSM.

According to the reentrant theory of Di Lollo et al. (2000), the behavioral performance is impaired if a mismatch occurs between a reentrant signal from a high-level area and an ongoing low-level representation. In the trailing mask condition, the high-level representations of mask and target are projected to the low-level area, which contains only the mask when these signals arrive. The resulting mismatch is resolved in favor of the mask representation, diminishing the target representation, and thus impairs the target recognition. On the other hand in the common offset condition, the low-level area contains a weak, decaying representation of each the mask and the target. The high-level representations of mask and target are again projected backwards, but now match with the low-level representations. Thus, the representation of the target is preserved, resulting in a high recognition accuracy. However, the proposed reentrant explanation arouses some very disagreeing opinions from other researchers (Francis and Hermens, 2002; Pilling and Gellatly, 2010; Põder, 2013, 2014), and contradicting replies from Di Lollo against them (Di Lollo et al., 2002; Jannati et al., 2013). Therefore, we will pick up this open question in our discussion.

The second explanation of OSM is the object updating hypothesis (Lleras and Moore, 2003; Moore and Lleras, 2005; Pilling and Gellatly, 2010) which assumes an update process of object representations during perception. In OSM, the behavioral performance is impaired because the target representation is transformed into the mask representation during the update process. This incorrect updating requires that the mask is spatially close to the target (Di Lollo et al., 2000), and has similar feature properties like motion (Lleras and Moore, 2003) or color (Moore and Lleras, 2005).

However, it is still unclear which theory truly accounts for OSM. This motivates us to propose visual attention as an alternative explanation of OSM. Classically, OSM is thought to be a new kind of masking paradigm and that it requires a separate theory (Di Lollo et al., 2000). Yet, given that the classical reentrant theory got many disagreeing comments and that it is still unclear which theory accounts for OSM, we suggest to drop the idea of needing a separate theory for OSM. Instead, we would like to propose that OSM is a normal visual attention phenomenon that can be explained by known mechanisms underlying visual attention.

Visual attention is a concept to focus perception and visual processing resources on a particular stimulus, typically important for the current task (Hamker, 2005a). Attention can be deployed to particular features or objects (feature-based attention), or to locations (spatial attention). This concept comes with a rich set of neurophysiological (Chap. 2.3) and psychophysical effects (Chap. 2.5). Some studies indicate a relationship between OSM and visual attention. Di Lollo et al. (2000) concluded that attention has to be spatially divided among all items for OSM because of two observations: Firstly, OSM requires a high number of distractors on the display *(set size effect)*, which prevents the target from capturing attention. Secondly, attending the target location with a spatial cue prevents OSM (Di Lollo et al., 2000). In their model CMOS, the set size effect is implemented by a delayed deploy of attention dependent on the number of items. Another study, Põder (2013), sees CMOS as a kind of attentional gating model because CMOS reads out the target representation for a behavioral response when attention is deployed. These two and several other publications (Dux et al., 2010; Enns and Di Lollo, 1997; Luiga and Bachmann, 2008) suggest a link to attention, supporting our idea that an attention model may be able to explain the complete OSM paradigm.

Relevant for our OSM explanation is the well-known attention paradigm of biased competition (Desimone and Duncan, 1995; Desimone, 1998). There, a target and a distractor are presented within the receptive field of a neuron. If attention is directed away, neurons preferring the target react lower as if the target would be presented alone. If the target is attended, the neuronal response will increase, whereas if the distractor is attended, the response will decrease. Hence, attention has the effect to amplify neurons preferring an attended stimulus and to suppress neurons preferred an unattended stimulus (Chap. 3). Another relevant attention paradigm for our OSM explanation is guided visual search, in which a target object has to be searched among distracting objects (Wolfe, 1994; Lee and McPeek, 2013). There, attention is deployed to potential target objects until the target is found (Wolfe, 1994, 2000; Woodman and Luck, 1999).

The proposed model was developed in the last chapters to simulate attentive dynamics at neurophysiological level (Chap. 3) and in guided visual search (Chap. 4, there termed as object localization). Our model stems from a line of earlier attention models (Hamker, 2005a,b; Zirnsak et al., 2011a) that account among others things for the neurophysiological and psychophysical aspects of visual search. All such models are composed of some of the major areas involved in attentional control (Miller and Buschman, 2013): mid- and high-level visual areas like V4 or IT; areas involved in spatial attention like the frontal eye field (FEF); and the lateral prefrontal cortex (PFC) which controls attention for the purpose of the current task (Koechlin, 2003; Sakai, 2008). As such models contain multiple brain areas, we denote them as system-level models. System-level models of attention can explain many psychophysical phenomena, an overview is given in Chap. 2.5.

The remaining of this chapter is organized as follows. Firstly, we will describe the novel model of attention (Sec. 5.2), and afterwards investigate OSM from three perspectives with the help of our model (Sec. 5.3). Firstly, we will show that the model can fit the behavioral data. Secondly, we will explain OSM by illustrating its temporal dynamics. Thirdly, we will link these explanations to the underlying attention mechanisms, and in this way, explain OSM as attention phenomenon. In the following section 5.4, we will discuss our elucidations in relation to existing OSM theories, show which aspects of these existing theories are consistent with our model, and address open research questions in OSM. Finally in section 5.5, we will conclude that OSM can be indeed explained by a model of visual attention and illustrate future research directions.

5.2. Model

The model was developed in the last chapters to model the neurophysiological properties of visual attention (Chap. 3) and the attentive dynamics in guided visual search (Chap. 4). Now, we simulate with it OSM to show that a model of attention can fully account for OSM. We will firstly give an overview of our model (the full description can be found in Chap. 4.2), before we will outline its modifications and relevant aspects for OSM.

5.2.1. Overview

The model (Fig. 5.1) contains a lower visual area (LGN/V1) and a higher visual area (HVA). The lower visual area models the lateral geniculate nucleus (LGN) plus the primary visual cortex (V1), and contains cells encoding color-contrasts, edges, or single dots. The higher visual area simulates cortical areas like V4 or IT. In OSM, HVA yields the neuronal representation of target, distractors, and mask. Each one is distributively represented by multiple neurons (Bridgeman, 2007), denoted as features. OSM utilizes as target stimulus a ring with a gap at one side (Landolt ring, Danilova and Bondarko (2007)), which is represented by a neuron preferring the ring and by a neuron preferring the gap. Distractors are other randomly-oriented Landolt rings, and are identically represented. The mask is represented in a similar way by two neuronal features. The visual areas are complemented

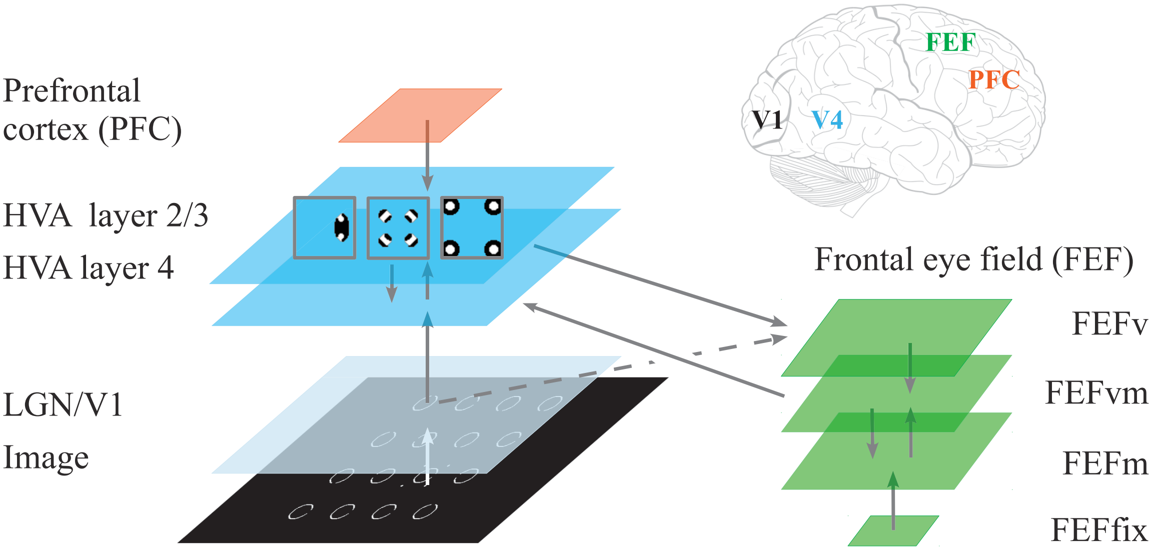


Figure 5.1.: The system-level model of visual attention in OSM. The model consists of a lower visual area (LGN/V1) encoding edges or single dots, a higher visual area (HVA) encoding distributively the rings and the four-dot mask, and the prefrontal cortex (PFC) projecting top-down attention signals to HVA. Those cortices are complemented with a model of the frontal eye field (FEF), split into: visual (FEFv), visuomovement (FEFvm), movement (FEFm), and fixation cell types (FEFfix). The dotted line denotes a fast pathway from LGN/V1 to FEF, simulating the fast dorsal pathway LGN→MT→FEF in the cortex. The model is described in Chap. 4 and is slightly modified here for OSM.

with a model of the prefrontal cortex (PFC) encoding object or task related information, and a model of the frontal eye field (FEF) simulating spatial processing as well as gaze control. These areas are all prominently involved in attentional processing (Hamker, 2005a; Miller and Buschman, 2013).

Attentional processing is realized in the model via a set of neuronal mechanisms: 1) amplification of neurons preferring an attended stimulus, 2) divisive normalization, 3) spatial pooling within the receptive field of a neuron, and 4) suppression between neurons. The mechanisms are provided by the microcircuit model of attention (Chap. 3), which simulates here area HVA. Due to that model, this area is composed of two layers, a modulatory layer 4 and a spatial pooling layer 2/3. The microcircuit model explains a very large amount of neurophysiological attention effects with the four neuronal mechanisms, hence they are solidly grounded on large corpus of attentional data. As attention can be deployed to a particular feature or location, the model contains a feature-based and spatial variant of the amplification as well as of the suppression mechanism. Feature-based attention is assumed to amplify the response of particular features in both HVA layers (feature-based amplification mechanism). This is implemented via the chain PFC→HVA layer 2/3→HVA layer 4, modeling the known cortical top-down control of attention (Miller and Buschman, 2013). It is accompanied with local inhibitory connections from HVA layer 2/3 to layer 4, suppressing neurons encoding a different feature (feature-based suppression). Spatial attention is assumed to emerge via recurrent processing between higher visual areas, e.g. HVA, and spatial maps like the FEF (Hamker, 2005a). In the model, this results in an amplification signal from FEFvm to HVA layer 4 towards the attended location (spatial amplification), and in a suppressive signal towards all other locations (spatial suppression).

The frontal eye field (FEF, Pouget et al. (2009)) is a brain structure involved in spatial processing and gaze control. The model is based on Zirnsak et al. (2011a) and consists of the following cell types (Bruce and Goldberg, 1985; Schall, 1991): visual cells reacting only to visual stimuli (FEFv); movement cells encoding the preparation of a saccade towards a specific target (FEFm); visuomovement cells responding to both influences (FEFvm); and a fixation cell (FEFfix) suppressing the execution of saccades (Hamker, 2005a; Hasegawa et al., 2004). The FEF realizes a competition between locations to focus activity on a single location, designating the target. This competition is implemented via local excitatory connections and long-range inhibition from FEFv to FEFvm. With many distractors, this competitive processing takes more time to focus activity on a single location. This effect accounts in predecessor models for the increased reaction times in visual search with many items (Hamker, 2005a), and accounts here for the set size effect (Sec. 5.3.3). The FEF projects back to HVA, forming a recurrent loop, which focuses neuronal activity on a single location and leads to the emergence of spatial attention (Hamker, 2005a). In OSM, this is covert spatial attention as OSM setups typically do not allow eye movements. The FEF receives input not only from HVA, but also via a fast pathway from LGN/V1. This pathway models the fast dorsal pathway LGN→MT→FEF in the cortex (Heinzle, 2006; Sincich et al., 2004). The stimulus signal reaches the FEF via it faster than through HVA (70 ms vs. 110 ms), which biases the FEF activity for a salient or a pop-out target before information from HVA arrive. The pop-out bias explains the high recognition accuracy with small set sizes (Sec. 5.3.3). The pathway was disabled in the object localization task (Chap. 4) as it does not play a relevant role there.

In summary, our attention model is strongly rooted on physiological data via the mechanistic microcircuit model of attention and the modeled cell types of the frontal eye field. Moreover, we ground our model on a large corpus of neuroanatomical data, like cortical connectivity and areas, cortical control via attention, receptive field sizes, and synaptic transmission delays. The two latter will be depicted in the next section.

5.2.2. Relevant model aspects and modifications for OSM

The model is described by the same set of equations as in Chap. 4, thus we would refer the reader to this chapter. Here, we will elaborate only model aspects particular relevant for OSM. This encloses the neuroanatomical aspects of the model like receptive field sizes or accurate synaptic transmission delays, which we explain here instead of Chap. 4 because they are more relevant for the correct modeling of a psychophysical experiment like OSM than for an applied-oriented task like object localization. Furthermore, we will present model modifications required for OSM, like learning of neuronal representations, maintaining of neuronal responses, or behavioral decisions. Finally, we will list the chosen model parameters.

Receptive field sizes

The model contains realistic receptive field (RF) sizes to intensify its neuroanatomical foundation and to allow an accurate modeling of psychophysical experiments. The sizes are user-definable for each area, whereby the standard values are 0.5° for LGN/V1, 3° for HVA layer 4, and 5° for HVA layer 2/3. For simplicity, the receptive field sizes are constant across eccentricity. The standard sizes are defined according to the reported data of Smith et al. (2001, Fig. 9), at an arbitrary chosen eccentricity of 5°. The receptive field sizes of HVA are based on area V4, whereby we choose layer 2/3 cells according to the physiologically reported data and layer 4 cells to some degree smaller as such cells have smaller fields (Chap. 3). LGN/V1 filters the image only at a single spatial frequency, however the user can fit this frequency to the stimulus material by changing the receptive field size. The available range is 0.033° (or 2') to 0.5°, as reported for LGN and V1 by Dow et al. (1981, Fig. 7, 9), by Parker and Hawken (1988, Fig. 4, 5), and briefly by Hubel and Wiesel (1974) or by Wässle and Boycott (1991).

Connectivity	Delay in ms
Image → LGN/V1	60
LGN/V1 → HVA L4	40
HVA L2/3 → FEFv	10
FEFvm → HVA L4	10
HVA ↔ PFC	10
Image → LGN/V1 → FEFv	70
Within the areas HVA and FEF	1

Table 5.1: Synaptic transmission delays in the model.

In the current OSM experiment, the receptive field sizes were adapted to the setup of Di Lollo et al. (2000). We chose 0.1° for LGN/V1 and 1.3° for HVA layer 2/3. The first value is suitable to detect the dots and rings. The latter value was chosen as this size encloses the four-dot mask. This HVA size lies between the field sizes of V2 and V4 according to Smith et al. (2001), if we consider an eccentricity of 1.7° which is the average stimulus eccentricity in the setup. However, Smith et al. (2001) used a linear function to estimate the size at the eccentricity of 1.7°, whereas the recent study of Motter (2009) shows that V4 receptive field sizes follow the same cortical magnification law as V1. This indicates that V4 has smaller RFs at lower eccentricity, like 1.7°, than estimated by Smith et al. (2001), thus we anticipate that HVA correspond to V4 in terms of receptive field sizes.

Temporal dynamics

Important for masking paradigms are accurate temporal dynamics. Thus, we model realistic synaptic transmission delays for each connection (Tab. 5.1). The data of the feedforward connections is taken from Schmolesky et al. (1998, Fig. 2), and if a connectivity is not available there, from Bullier (2001, Fig. 1). We anticipate again that HVA correspond to V4. Delays within an area have been reported to be between 0.5 and 1.5 ms (Thomson et al., 2002, Tab. 1), thus we choose 1ms for all connections. No data is available for the connections between FEF and HVA, so we assume 10 ms similar to the connection between HVA and PFC.

Behavioral decisions

In OSM, the subjects have to recognize the target Landolt ring by identifying the gap in ring. The gap can be at four positions, thus the subjects have to decide between four alternatives. To simulate such a behavioral decision, we extend our in Chap. 4 described model with a decision process. We presume that the probability to identify the gap correctly depends on the integrated evidence for it. In our model, the gap is represented by the neuronal activity of a gap preferring neuron in HVA layer 2/3. Thus, we integrate this neuronal response over the entire simulation time (Eq. 5.1) and model a probabilistic decision process based on this value R. If this value is over a perceptual threshold θ, the correct decision is always chosen, and otherwise a random one. Four choices are available, thus the random decision will be correct with 25% probability too (Eq. 5.2). The threshold θ_x is drawn randomly for each trial x from a Gaussian distribution with μ_θ and σ_θ. The variable threshold models the behavioral variability of an unbiased ideal observer. Finally, the per-trial correctnesses C_x are averaged to the total correctness of the experiment C (Eq. 5.3). This latter value is also often denoted as recognition accuracy.

$$R_x = \sum_{t=0}^{t_{end}} r_{target}(t) \tag{5.1}$$

$$\text{Correctness of trial x: } C_x = \begin{cases} 1 & R_x > \theta_x \\ 1 \text{ with } P = 0.25 & R_x \leq \theta_x \\ 0 \text{ with } P = 0.75 & R_x \leq \theta_x \end{cases} \tag{5.2}$$

$$\text{Total correctness: } C = \bar{C_x} \tag{5.3}$$

Learning of neuronal mask and ring representations

The neuronal representations of the mask and the Landolt rings in HVA were created by a simple one-shot learning procedure (Eq. 5.4 - 5.5). The procedure calculates the connectivity matrix ($w^{\text{V1-HVA4}}$) between V1 and HVA layer 4 directly from the stimulus (Eq. 5.4), and scales the matrix in such a way that the HVA neurons are excited maximally ($E^{\text{HVA4}} = 1$) for their preferred stimulus (Eq. 5.5). The preferred stimulus is represented as a specific pattern of V1 complex cell activity, r^{V1C}. A stimulus is represented distributively by multiple feature cells, each cell encoding a different spatial part. The part is specified by a manually designed area selection matrix w^{part} with binary elements. The procedure learns the connectivity matrix offline for a single HVA location and shares it with all locations (weight sharing approach). Compared to the trace learning from the previous chapter, the one-shot learning approach is simpler, but also has less free parameters and is by far computationally faster, resulting in a much lower effort to learn the stimuli. As the OSM stimuli are not very complicated to learn, we choose this approach over the trace learning from the last chapter.

$$\begin{aligned} B_{i,\,d',i',x'} &= \left[P\left(r^{\text{V1C}}_{d',i',x'}\right) - \nu\right] \cdot w^{\text{part}}_{i,x'} & (5.4)\\ w^{\text{V1-HVA4}}_{i,\,d',i',x'} &= B_{i,\,d',i',x'} \,/ \sum_{d'',i'',x''} \left(B_{i,\,d'',i'',x''} \cdot r^{\text{V1C}}_{d'',i'',x''}\right) & (5.5) \end{aligned}$$

Whereby $r^{\text{V1C}}_{d',i',x'}$ denotes the response of a V1 complex cell in channel d', feature i', and location x'. The index i denotes the feature in HVA. Thus, the connectivity matrix has four indices: $w^{\text{V1-HVA4}}_{i,\,d',i',x'}$. $P(x) = x^2$ denotes a non-linearity to favor stronger weights and B an intermediate variable. The parameter ν controls the relative amount of inhibitory weights and is individually chosen for the ring and mask stimulus: $\nu_{ring} = 0.25$, $\nu_{mask} = 0.15$.

The following equations proof that a HVA layer 4 neuron is indeed excited (E^{HVA4}) with 1 by its preferred stimulus r^{V1C}. For better readability, the presynaptic indices $\{d', i', x'\}$ are grouped together into a single presynaptic index j'.

$$\begin{aligned} E^{\text{HVA4}}_i &= \sum_{j'} \left(w^{\text{V1-HVA4}}_{i,\,j'} \cdot r^{\text{V1C}}_{j'}\right), \quad \text{with: } w^{\text{V1-HVA4}}_{i,\,j'} = B_{i,\,j'} \,/ \sum_{j''} \left(B_{i,\,j''} \cdot r^{\text{V1C}}_{j''}\right)\\ &= \sum_{j'} \left([B_{i,\,j'} \,/ \sum_{j''} \left(B_{i,\,j''}\, r^{\text{V1C}}_{j''}\right)] \cdot r^{\text{V1C}}_{j'} \right)\\ &= \sum_{j'} \left(B_{i,\,j'}\, r^{\text{V1C}}_{j'} \,/ \sum_{j''} \left(B_{i,\,j''}\, r^{\text{V1C}}_{j''}\right) \right)\\ &= \sum_{j'} \left(B_{i,\,j'}\, r^{\text{V1C}}_{j'}\right) \,/ \sum_{j''} \left(B_{i,\,j''}\, r^{\text{V1C}}_{j''}\right)\\ &= 1 & (5.6) \end{aligned}$$

Maintaining of neuronal responses

A requirement for OSM is to maintain the neuronal representation of a stimulus after its offset. We use for this purpose spatiotemporal receptive fields in V1 (DeAngelis et al., 1993; Cai et al., 1997) eliciting a swift decaying response after stimulus offset, and attention to maintain this response (Sec. 5.3.3).

The temporal characteristic of a spatiotemporal receptive field is a strong transient peak-response to a stimulus onset and a decaying response afterwards (for example illustrated in Fig. 5.8 in Sec. 5.3.3). We model spatiotemporal fields via separated spatial and temporal filters. The spatial filter is already implemented in the LGN/V1 model described in Chap. 4, producing a static response r^{V1S}. The temporal characteristic is implemented on top of that via an alpha and a z-function. These functions have been developed within the group of Prof. Hamker (Appendix C) and are refined here to operate with continuous temporal input. The z-function immediately peaks at stimulus onset (Eq. 5.7) and moderately declines afterwards (Eq. 5.8). The alpha-function follows this z-function, resulting in a swift increase plus a slow decay (Eq. 5.9). Transient responses typically decline to a baseline greater than zero (Fischer and Boch, 1985), hence we add a baseline (Eq. 5.10). After stimulus offset, the response will decay slowly from baseline to zero, simulated by another ODE (Eq. 5.11). The responses of the resulting spatiotemporal V1 simple cells are denoted as r^{V1T}. The responses will be projected to the next layer, V1 complex, as in the original model (Chap. 4).

$$\begin{aligned}
z_{d,i,x} &= \max\left(\left[r^{\text{V1S}}(t) - r^{\text{V1S}}(t-1)\right]^{+}, z\right) && (5.7)\\
\tau_z \frac{\partial z_{d,i,x}}{\partial t} &= -z && (5.8)\\
\tau_{imp} \frac{\partial p_{d,i,x}}{\partial t} &= -p + z && (5.9)\\
q_{d,i,x} &= v^{\text{imp}}\, p + v^{\text{base}}\, r^{\text{V1S}} && (5.10)\\
\tau \frac{\partial r^{\text{V1T}}_{d,i,x}}{\partial t} &= -r + q \quad \text{with:} \quad \tau = \begin{cases} 1 & \partial r \geq 0 \\ \tau_{trace} & \partial r < 0 \end{cases} && (5.11)
\end{aligned}$$

Whereby $\tau_z = 5$ denotes the time constant for the z-function, $\tau_{imp} = 20$ for the alpha function, and $\tau_{trace} = 40$ for the slow decline of the response. The proportion of baseline activity is defined via the parameter $v^{\text{base}} = 0.25$. To ensure that the peak amplitude of the response is preserved between r^{V1S} and r^{V1T}, we use a scaling parameter $v^{\text{imp}} = 7.7$.

Dorsal pathway and surround suppression

The OSM setup uses two further aspects that were not necessary in the object localization setup and hence were deactivated there (Chap. 4): the dorsal pathway LGN/V1→FEF and surround suppression. Both were activated by setting the scaling parameter of the pathway ($v^{\text{V1-FEFv}}$, Tab. 5.2) and of the suppression ($v^{\text{SUR-1}}$, Tab. 5.2) to positive values.

The former is involved in the set size effect (Sec. 5.3.3), while the latter is necessary to suppress incorrect HVA responses towards the 4-dot mask pattern. HVA layer 4 neurons encoding the four-dot mask respond correctly in the center of the mask stimulus, whereby each dot contributes with 25% to the excitation. However, such a single dot falls also in the receptive fields of adjacent HVA neurons in the surround and elicits in them an incorrect response. Single cell recordings show a suppression between such center and surround neurons (surround suppression mechanism, Chap. 3.3.4). In OSM, the center neuron is highly active, thus this mechanism inhibits the incorrectly responding surround neurons.

Parameter set

We now list the parameter values used in OSM (Tab. 5.2) and discuss them. In the OSM setup, we would like to remain closely on the parameter values used for fitting the neurophysiological data (Chap. 3) to achieve a solid physiological foundation of the model. For this, we set the amount of feature-based amplification to $v^{\text{PFC-HVA2}} = 0.5$. This value corresponds to the microcircuit model fitting the physiological data (Chap. 3). In the object localization scenario (Chap. 4), it was higher as the model was optimized for highest localization performance.

Besides this, some other parameter values differ to Chap. 3 and 4 as the excitation in HVA layer 4 and the size of the neuronal areas differ. The excitation is much sparser in the OSM setup than in the microcircuit model or in the object localization setup. The microcircuit model uses a non-sparse input in form of broad symmetrical exponential functions, while in object localization, the background causes a broad excitation of HVA. To compensate this, we strengthen the signal from layer 4 to layer 2/3 by increasing $v^{\text{HVA4-HVA2}}$ from 1 to 8.

In OSM, the spatial dimensions of HVA layer 4 and the FEF are five-times larger than in the object localization task. This influences spatial and surround suppression as they are driven by more presynaptic neurons. To account for this, we increase the non-linearity parameter of spatial suppression from $p^{\text{SP-2}} = 1$ to 4, and of the surround from $p^{\text{SUR-1}} = 1$ to 2 and from $p^{\text{SUR-2}} = 2$ to 20. The scaling values were also adapted appropriate to the changed non-linearities.

Description	Value
V1	
LGN and V1 receptive fields	$size = 0.1°$
Standard deviations of DoGs	$\sigma_{\text{L/M-c}} = 1.5$, $\sigma_{\text{L/M-s}} = 6$, $\sigma_{\text{LM/S}} = 6$
Standard deviations of Gabors	$\sigma_1 = 4.5$, $\sigma_2 = 18$
V1 complex	$p_{V1C} = 1$
HVA layer 4	
Excitation	$v^{\text{V1-HVA4}} = 1.5$, $p^{\text{E}} = 1$
Cell response	$\sigma = 0.3$, $g^{\text{HVA4}} = 1.05$, $\tau^{\text{HVA4}} = 10$
Feature-based amplification	$v^{\text{HVA2-HVA4}} = 1$, $p^{\text{HVA2-HVA4}} = 1$
Spatial amplification	$v^{\text{FEFvm-HVA4}} = 4$
Feature-based suppression	$v^{\text{FEAT-1}} = 2.2$, $p^{\text{FEAT-1}} = 3$, $v^{\text{FEAT-2}} = 1.35$, $p^{\text{FEAT-2}} = 4$
Spatial suppression	$v^{\text{SP-1}} = 0.002$, $p^{\text{SP-1}} = 1$, $v^{\text{SP-2}} = 4$, $p^{\text{SP-2}} = 4$
Surround suppression	$v^{\text{SUR-1}} = 1.5$, $p^{\text{SUR-1}} = 2$, $v^{\text{SUR-2}} = 1.75$, $p^{\text{SUR-2}} = 20$
HVA layer 2/3	
Excitation	$v^{\text{HVA4-HVA2}} = 8$, $P_1 = 4$, $P_2 = 0.25$
Cell response	$\sigma = 2$, $g^{\text{HVA4}} = 1.22$, $\tau^{\text{HVA2}} = 10$
Feature-based amplification	$v^{\text{PFC-HVA2}} = 0.5$
FEFv	
Excitation	$v^{\text{HVA2-FEFv}} = 1$, $v^{\text{V1-FEFv}} = 0.125$
Cell response	$\tau^{\text{FEFv}} = 10$
Competition	$c = 1.8$
Signal enhancement	$\sigma = 0.1$
FEFvm	
Excitation	$v^{\text{Ev}} = 0.2$
Cell response	$\tau^{\text{FEFvm}} = 10$
Suppression	$v^{\text{Sv-1}} = 0.18$, $p^{\text{Sv-1}} = 0.4$, $v^{\text{Sv-2}} = 0.35$
FEFm	
Excitation	$v^{\text{FEFvm-m}} = 1.3$
Cell response	$\tau^{\text{FEFm}} = 10$
Suppression	$v^{\text{Svm}} = 0.3$, $v^{\text{Sfix}} = 3$
Saccade threshold	$\Gamma^{\text{FEFm}} = 0.9$

Table 5.2.: Model parameters in object substitution masking. The parameters refer to the model description in Chap. 4.

5.3. Results

In this section, we will investigate OSM via three perspectives. We will first demonstrate that the attention model can account for OSM by fitting the behavioral data of two OSM experiments, the original study of Di Lollo et al. (2000) and a recent study from Argyropoulos et al. (2013). Secondly, we will explain OSM with the help of our attention model, and thirdly, we will link these explanations to the underlying neuronal attention mechanisms, realizing suppression, maintaining of responses, and spatial amplification.

5.3.1. Behavioral data and model fits

Study of Di Lollo et al. (2000)

The study of Di Lollo et al. (2000, experiment 3) was the first reporting OSM. Their setup (Fig. 5.2a - c) consists of up to 16 rings with a gap, denoted as Landolt rings (Danilova and Bondarko, 2007). One of the rings is indicated as the target by four surrounding dots. The task of the subjects is to identify the position of the gap. The setup contains four different gap-positions (top, bottom, left, right), thus chance level of a correct recognition is 25%. The same four-dot pattern serves also as the mask and is set off a variable time span after stimulus offset (0 ms - 320 ms). The experiment consists of 25 different conditions in which the mask duration and number of rings (set size) were systematically varied.

In addition to mask duration and set size, Di Lollo et al. (2000) varied within the experiment the spatial distance between target and mask as well as the target location. They observed no significant effect for altering the distance, thus we will not model this variation within the experiment. The variation of the target location showed an eccentricity dependent masking effect, whereby the shapes of the performance curves were similar for all eccentricities. However, they did not report the different conditions. Instead, they concluded that an average over all eccentricities would still show the same curve shape, and it is sufficient to report only the averaged curves. Hence, the eccentricity seems to be unimportant, so we do not model it. This also implies that we do not model receptive field size variations caused by eccentricity. We choose instead a single eccentricity and model the receptive fields according to the known physiological sizes at this eccentricity (Sec. 5.2). As eccentricity, we choose the average eccentricity of all items. Di Lollo et al.

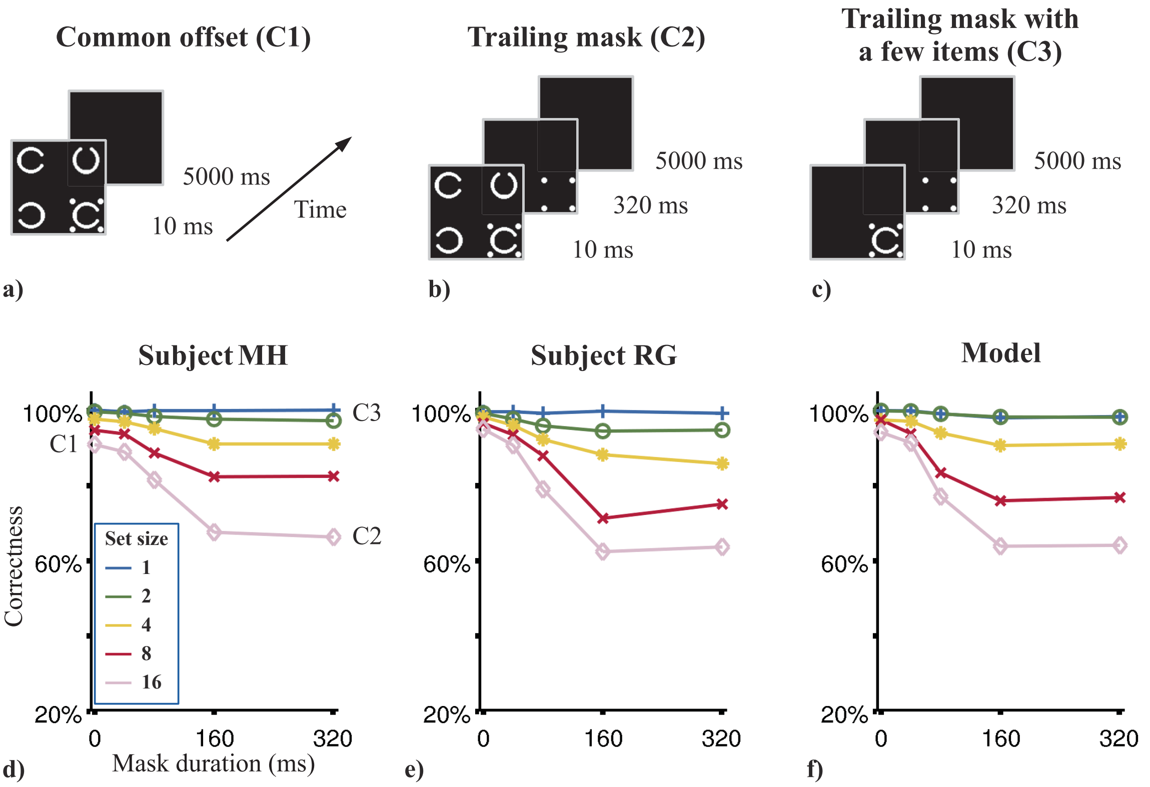


Figure 5.2.: OSM study of Di Lollo et al. (2000). **a-c)** Experimental setup which consists of a target Landolt ring indicated by four dots and multiple other rings serving as distractors. The four-dot pattern serves as the mask too. **a)** Common offset condition C1: Target and mask are switched off at the same time. The condition uses 16 items, from which four are illustrated. **b)** Trailing mask condition C2: Like C1, but the mask lasts longer than the target, resulting in an impaired behavioral performance. **c)** Trailing mask condition C3: Like C2, but only with one item. Contrary, this results in an unimpaired behavioral performance, denoted as set size effect. **d, e)** Behavioral data of Di Lollo et al. (2000). Within the experiment, the mask duration and the set size were systematically varied. The conditions C1 - C3 denote the extreme cases in the experiment. **f)** Model fit of the data.

(2000) reported that 4 items were placed at 0.7° eccentricity, 8 items at 1.6°, and 4 items at 2.8°, resulting in an average eccentricity of 1.675°.

Di Lollo et al. (2000) calibrated the stimulus luminance so that they appear equally bright to the human observers. We presume that this results in an equally strong input signal in the higher areas of the visual cortex, and thus ensure that the mask and target stimuli excite HVA with the same value, here arbitrary chosen as 1.

The attention model is fully deterministic except the decision process, thus we do not need to repeat the experiment several times as done with human subjects. Contrary, the decision process is probabilistic as it simulates the varying decision of human observers. For this process, we run 1600 trials for each of the 25 conditions. This amount of repetitions was also used by Di Lollo et al. (2000).

We will focus on the extreme cases of the 25 conditions, denoted as *C1*, *C2*, and *C3*. The *condition C1* represents a common offset condition in which the mask is synchronously set off with the stimulus (Fig. 5.2a). We consider a set size of 16 items in C1, as well as in C2. Despite the short presentation time of 10 ms, the target is recognizable, demonstrated by the 90% and 95% correct response of the subjects MH and RG (Fig. 5.2d, e). However, if the mask is kept after target offset in the same setup (trailing mask condition), the behavioral performance is significantly impaired. We investigate this behavior in *condition C2* (Fig. 5.2b), using the maximum mask duration of 320 ms and the number of items with the highest performance impairment, 16 items. The performance begins to drop at 40 ms mask duration until it reaches it minimum value of about 65% at 160 ms mask duration. Interestingly, the performance is only impaired with a high number of items, denoted as set size effect. Thus, we consider a trailing mask condition with a few items too. We based this *condition C3* on the condition C2, except the display contains only one item (Fig. 5.2c).

The experiment of Di Lollo et al. (2000) was obtained with only two subjects from whom the individual data sets are shown (Fig. 5.2d, e). The model was fitted as a compromise between both data sets. It shows all three significant effects found by Di Lollo et al. (2000) in OSM (see Fig. 5.2f): a high behavioral performance in all common offset conditions (e.g. C1), an impaired performance in trailing mask conditions with multiple distracting items (C1 versus C2), and no impairment with few items (C2 versus C3).

Study of Argyropoulos et al. (2013)

The study of Argyropoulos et al. (2013) investigates primarily the interaction of set size and mask duration. They found that the two variables are indeed independent, a finding which we will discuss deeply in Sec. 5.4.2. However, much more interesting for modeling purpose is that they extended and refined the setup of Di Lollo et al. (2000). They obtained more reliable data by a) using ten subjects instead of only two, b) placing all items on the same eccentricity, and c) altering the task difficulty to avoid ceiling or floor effects in the performance curves. The latter was accomplished by calibrating the eccentricity and gap

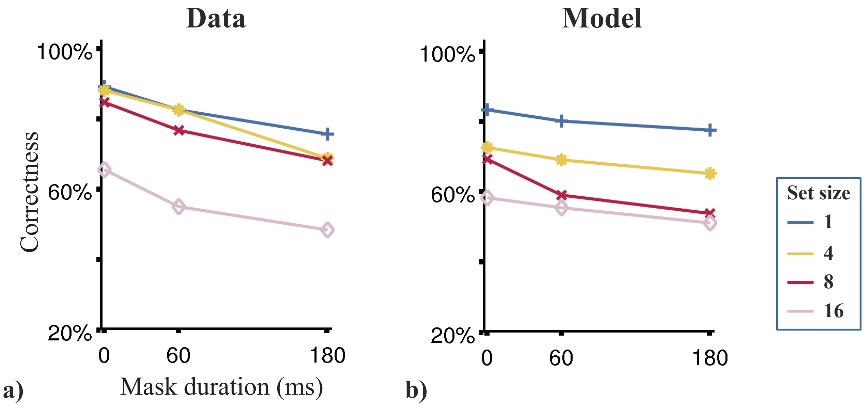


Figure 5.3.: Data of Argyropoulos et al. (2013, experiment 3) **(a)**, and fits by the proposed model **(b)**. The data shows the averaged performance over ten subjects. Compared to the experiment of Di Lollo et al. (2000) (Fig. 5.2), long mask durations like 320 ms were not tested. In consequence, the curves do not show the saturation effect occurring for higher mask durations than 160 ms.

size so that the maximal performance was below 100% and the minimum performance over chance level (again at 25%). Most interestingly, they found a masking effect also for one item (Fig. 5.3a). This effect was probably not observed by Di Lollo et al. (2000) due to a ceiling effect, i.e. because the performance in their experiment was in all one-item conditions at 100%.

Here, we simulate their findings in principle. We use the stimuli of Di Lollo et al. (2000), and increase the task difficulty by shrinking the gap of the target ring. In our model, this decreases the input signal into HVA cells representing the gap, and thus diminishes this target feature. The data was best fitted with a gap size that excites HVA at 60% of its maximum value. Besides the changed input signal, the model is identically for both data sets.

The model fit (Fig. 5.3b) replicates all aspects of data that have been reported as significant by Argyropoulos et al. (2013): These are a) a masking effect at all set sizes, even for one item, and b) the normal set size effect, i.e. that the behavioral performance is lower with more items. The slopes of the performance curves are slightly lower in our model, however the significance of this effect was not investigated. Furthermore, we observed that the performance curves fit well for 1 and 16 items, but not for 4 and 8 items. Yet, the study's principal investigator, Dr. Pilling, stated himself that our model shows very reasonable curves also for 4 and 8 items (personal communication at ECVP 2014), as their own data was rather unexpected. Unfortunately, they did not perform any significance test to check if this was a relevant effect or attributed to noise, as this aspect of the data was not the focus of their study (their focus was the novel masking effect with one item). Dr. Pilling personally suspected noise at the ECVP 2014, because the variance between subjects was quite high in the study. As our model replicates all significant effects of the study, especially the newly-found masking effect with set size one, and this data aspect was not checked for any significance, we think our model fits the data reasonable well.

5.3.2. Explanation of OSM

We will now explain the masking effect in OSM, i.e. the impairment of the behavioral performance, by comparing the baseline condition C1 with the condition C2 showing the effect. As the masking effect occurs only with a high number of items like in C2 (set size effect), we will contrast C2 with the one item condition C3 showing no masking effect. In each condition, we will illustrate the temporal processing in our model.

Baseline condition C1

The baseline condition C1 depicts a common offset condition, i.e. target and mask are set off simultaneously, with 16 items. According to our model, the following neuronal dynamics occur in this condition (Fig. 5.4): At 0 ms, target and mask stimulus are set on. Their signals reach HVA layer 2/3 at 100 ms and elicit strong responses in target as well as mask encoding neurons. Both stimuli are set off simultaneously 10 ms later, and their responses are maintained in HVA, i.e. they decay slowly but not immediately. The responses are maintained via an attention mechanism as described in Sec. 5.3.3. The mask response is slightly higher than the target response due to feature-based amplification from PFC. This signal results from a moderate feature-based attention toward the mask, representing the task instructions. The OSM task requires at the beginning a search for the four dots in the display, thus it is a guided visual search task. Such a task is realized by feature-based attention to the searched object (Chap. 4), i.e. the mask. Suppression occurs between mask and target neurons in HVA. The suppression is an attention mechanism and will be described in Sec. 5.3.3. As the mask response is only slightly higher than the target, the suppression towards the target is only slightly stronger than towards mask. Thus, both responses are suppressed: the target moderately and the mask weakly. These dynamics lead to a high mask and to an average high target response.

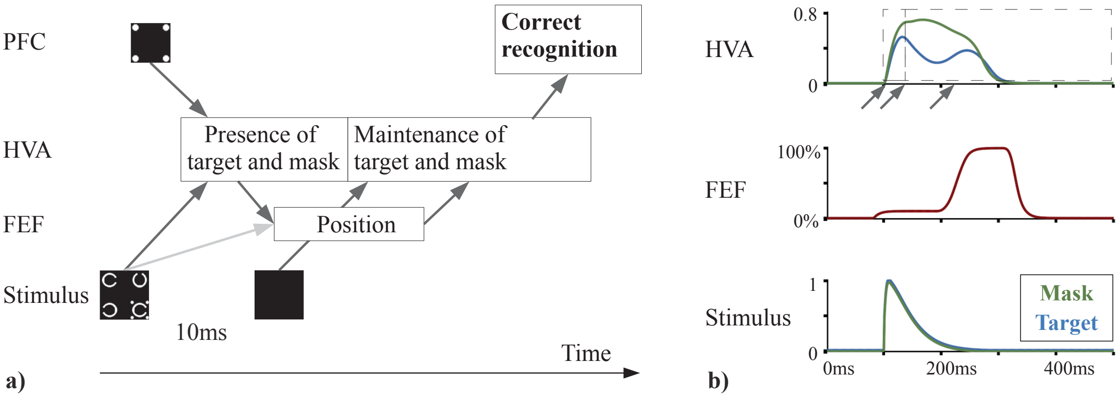


Figure 5.4.: Explanation of the baseline condition C1, common offset with 16 items. **a)** Schematic sketch of the temporal processing and signal flows between stimulus, FEF, HVA layer 2/3, and PFC. HVA layer 2/3 contains the neuronal representation of the target and the mask stimulus. The FEF encodes spatial information, resulting in amplification of the target position in HVA and in suppression of all other positions. The recurrent processing between HVA and FEF occurs permanently, however we only display two arrows for clarity. **b)** Modeled responses and signals from a). The recognition accuracy depends on the integrated target signal (blue), represented by a HVA layer 2/3 neuron encoding the target gap. Therefore, its response is shown along with all responses and signals influencing it: the mask response in HVA layer 2/3, the spatial amplification signal from FEF, and the stimulus at the target location. The signal from the FEF represents the amount of spatial amplification received at the target location by the HVA cells encoding target and mask. The stimulus is represented by the received excitation from LGN/V1 to HVA layer 4. The arrows illustrate the in a) shown signals from stimulus and FEF towards HVA: 1. onset of the stimulus, 2. offset of the stimulus, 3. onset of spatial amplification. The two HVA phases in a) are illustrated by dotted boxes.

Spatial attention to the target location emerges from the recurrent processing between HVA and FEF. At neuronal level, this attention results in an amplification of the target location and suppression of all others. Spatial attention emerges very late in C1 as the FEF needs some time with 16 items to focus neuronal activity on a single location (Sec. 5.2.1). Hence, the spatial amplification and suppression signals occur very late, typically around 220 ms after stimulus onset (Fig. 5.4b). At this time, the target response is amplified by spatial amplification. Overall, the target response is moderately strong, leading to 95% correctness in the setup of Di Lollo et al. (2000). In the much harder setup of Argyropoulos et al. (2013), the target signal has a similar temporal course, but it is weaker excited, leading to 70% correctness.

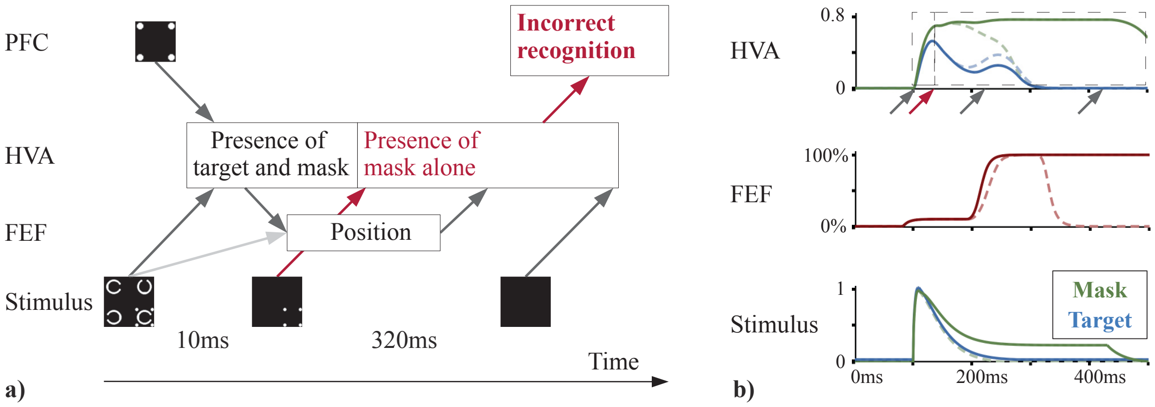


Figure 5.5.: Explanation of the condition C2, 320 ms trailing mask with 16 items (similarly illustrated as Fig. 5.4). The behavioral performance is impair in this condition C2 (masking effect), but not in the baseline condition C1 (Fig. 5.4). Conditions C2 and C1 are identical except for the trailing mask. **a)** Schematic sketch of the temporal processing. The differences because of the trailing mask are illustrated in red. **b)** Modeled responses and signals from a). Solid lines illustrate the current condition C2, whereas dotted lines show C1 for comparison. The performance impairment in C2 is caused by the more decreased HVA target response compared to C1 (solid versus dotted blue lines in HVA).

Masking effect illustrated by condition C2

The masking effect, i.e. the impairment of the behavior, occurs in the trailing mask condition C2 with 16 items. This condition is identical to condition C1, but the mask is shown 320 ms longer than the target stimulus. By this, the behavioral performance drops from 95% to 65% in the study of Di Lollo et al. (2000), and from 70% to 50% in the study of Argyropoulos et al. (2013). In our model, these drops are explained by the weaker target signal in HVA layer 2/3 (Fig. 5.5). At the beginning, the responses of target and mask are identical to condition C1, thus the mask response is high and the target response is average high. However due to the permanently presenting of the mask, the mask response remains high in C2 and becomes over time significantly stronger than the target. Due to this strong mask response, the suppression towards the target remains strong, resulting in a weak target response. When the spatial amplification signal from FEF arises at around 220 ms after stimulus onset, the target response is too low to be notable amplified and remains weak. Thus, the target response is especially after 220 ms weaker than in C1. In summary, the masking effect occurs because the target response is strongly suppressed by the mask response. Increasing mask durations lead to a longer highly-active mask neuron and thus to more suppression, resulting in a lower recognition accuracy.

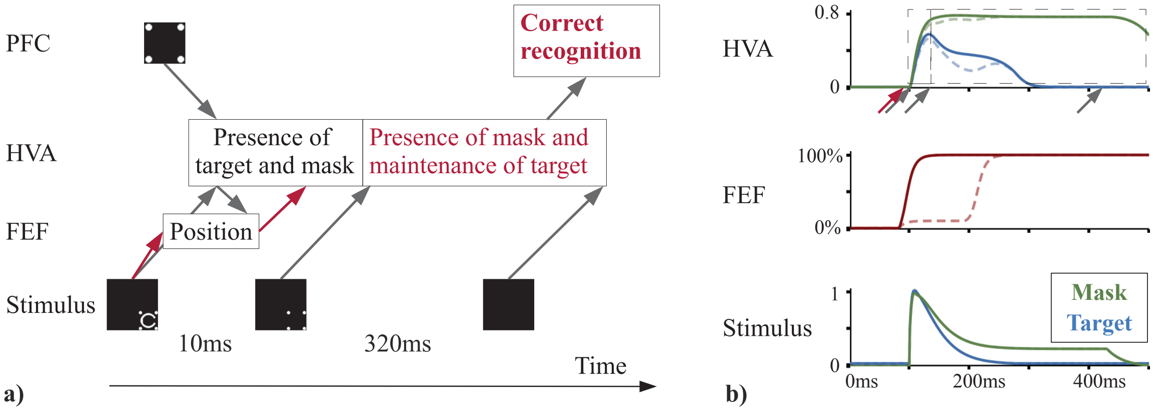


Figure 5.6.: Explanation of the condition C3, 320 ms trailing mask with one item. This condition does not show a masking effect like condition C2 (Fig. 5.5), demonstrating that multiple items are required for OSM (set size effect). The figure is identically illustrated to Fig. 5.5. **a)** Schematic sketch of the temporal processing. Differences to C2 are shown in red. **b)** Modeled responses and signals from a), shown as solid lines. For comparison, the responses and signals of condition C2 are shown as dotted lines. The high recognition accuracy is caused by the stronger HVA target response in C3 compared to C2 (solid versus dotted blue line in HVA).

Set size effect illustrated by condition C3

Condition C3 shows the set size effect in OSM. C3 is identical to the trailing mask condition C2, except the display contains only one item. In such a condition with a few items, the target can be recognized despite the trailing mask. The studies show this effect by maximal correctness rates in C3 and low rates in C2: 100% versus 70% in Di Lollo et al. (2000), and 80% versus 50% in Argyropoulos et al. (2013).

The reason is an early emergence of spatial attention in C3, which amplifies the neuronal representation of the target (Fig. 5.6). This underlying spatial amplification constitutes an attention mechanism and will be described in Sec. 5.3.3. The early emergence of spatial attention results from two facts. Firstly, the spatial processing converges faster with a few items to a single location. This relies on the competition between locations in the FEF, and this competition is faster with fewer items (Sec. 5.2.1).

Secondly, the model contains the fast cortical pathway LGN→MT→FEF (Sincich et al., 2004; Heinzle, 2006). In our model, this pathway evokes much earlier (at 70 ms) a FEF response on the item positions than via the normal pathway over HVA (at 110 ms). With one item, this immediately indicates the target position. Even with a few items, it constrains competition processing on a few positions and so speeds up the process. The pathway is also active in the conditions C1 and C2 with 16 items. However, this large number of items results in a very broad excitation of the FEF, which does not improve the competitive processing. In the FEF, competition is implemented via inhibition from FEFv to FEFvm. The broad and equally-strong response in the FEFv leads to a strong and equally-strong inhibition at all FEFvm positions. Thus, no position is favored until a signal arrives from HVA. Therefore, the pathway has no benefits with many items.

To conclude, spatial attention to the target location emerges earlier and is much stronger than in C2. This results in an amplification of the target response (Fig. 5.6b) as well as the mask response. The mask response increases slightly and ceils, but the target response is massively increased compared to C2. As the suppression between mask and target relies on these differences, the amount of suppression is decreased too. Both effects together lead to a stronger target response, so the target can be better recognized.

5.3.3. Attention mechanisms in OSM

According to our model, OSM relies on three neuronal attention mechanisms: 1) suppression to account for the masking effect, 2) amplification to maintain the neuronal representation after stimulus offset, and 3) spatial amplification to account for the set size effect.

Masking effect due to feature-based suppression

The masking effect relies on suppression between mask and target selective HVA neurons, denoted as feature-based suppression mechanism. The mechanism is part of the mechanistic microcircuit model of attention (Chap. 3), simulating area HVA. The feature-based suppression is transported via connections from HVA layer 2/3 to layer 4 (blue arrows in Fig. 5.7a). In OSM, the target neurons are suppressed by the strong response of the mask neurons, resulting from the trailing of the mask and an amplification signal from PFC due to feature-based attention (Sec. 5.3.2). The suppression occurs only between the ring and the mask at the target location as the suppressive connections are spatially limited, here to the extent of a HVA layer 2/3 receptive field.

As the microcircuit model is able to explain this masking effect, it indicates that it is an attention effect as the circuit was originally developed to explain a large amount of attention effects at neuronal level (Chap. 3). Compared to these effects, the dynamics in

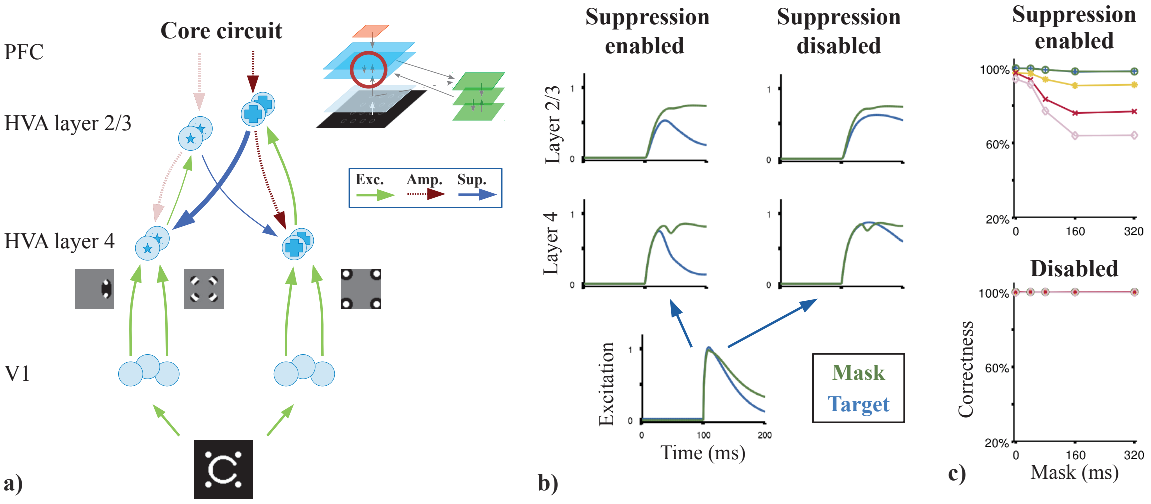


Figure 5.7.: The feature-based suppression mechanism accounts for the masking effect. **a)** The suppression mechanism (blue arrows) is part of the mechanistic microcircuit of attention, simulating HVA (red circle). HVA encodes distributively the target (left) and the mask (right). In OSM, the mask receives feature-base attention that amplifies neurons encoding the mask (red arrows). The connection type (excitation, amplification, suppression) describes the influence to its postsynaptic neuron. Attentional modulation of neurons and connections are denoted by the increase or decrease of the symbols thickness. **b)** Neuronal responses with and without the suppression mechanism. The figure shows in condition C2 the responses of HVA neurons encoding the target gap and the mask, plus the excitation towards HVA. **c)** Model performance when the suppression mechanism is disabled (bottom) in relation to the performance of the full model (top).

OSM are identical to the ones in the well-known attention paradigm of biased competition (Desimone and Duncan, 1995; Desimone, 1998). There, two stimuli suppress each other resulting in a competition. Attention can bias this competition for one of them by attending it, amplifying neurons encoding the attended stimulus and suppressing neurons encoding the unattended stimulus. These dynamics are directly transferable to OSM, as the mask selective neurons have an increased response due to feature-based attention and the trailing mask, and the target neurons are unattended and suppressed. This view enables us to explain the masking effect as biased competition paradigm.

The effect of suppression is illustrated by enabling versus disabling the mechanism in the model. Enabled, the responses differ swiftly, leading to the weak response of the target (Fig. 5.7b left). However if the mechanism is disabled, the responses are mostly similar whereby the mask is still higher due to amplification from PFC (Fig. 5.7b right). The disabling was achieved by cutting the suppressive connection between HVA layer 2/3 and

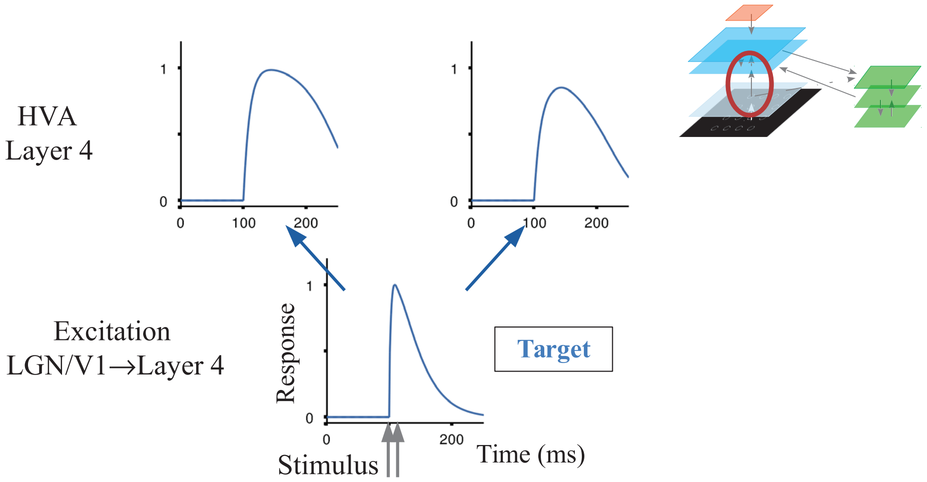


Figure 5.8.: The amplification mechanism maintains neuronal responses. The mechanism involves LGN/V1 and HVA layer 4, indicated by the red circle in the model. Attending the stimulus results in maintaining the neuronal responses after stimulus offset (left), whereby the response diminish faster without attention (right). Without attention, a weak maintenance also occurs due to divisive normalization in the model. The stimulus onset and offset are indicated by the two arrows (bottom).

layer 4 (blue arrows in Fig. 5.7a). To verify that the mechanism is indeed responsible for the masking effect, we check the behavioral performance when disabled. The result shows a perfect performance (Fig. 5.7c), proofing that the suppression mechanism accounts for the masking effect in OSM.

Maintaining of a neuronal representation via amplification

An important requirement for OSM is to maintain the neuronal representation of a stimulus after its offset. A short, transient response is often found in cortical neurons after stimulus offset and can be easily explained by the spatiotemporal structure of a neuron's receptive field. Yet, it is unclear which mechanism maintains such a transient response over a longer time period. We found out that attention can maintain such a signal as it amplifies the transient response and in this way effectively extends it.

Spatiotemporal receptive fields (Cai et al., 1997; DeAngelis et al., 1993) are found in the LGN, so we simulate them in the lower visual area (LGN/V1) of our model. In our model, we simulate their temporal characteristic of producing a sharp increasing and then decreasing response towards a stimulus. If the stimulus is set off, the response does not diminish immediately, instead it decreases swiftly to zero (Fig. 5.8 bottom). If the stimulus is presented permanently, the response decays to a baseline activity.

Attention amplifies this transient LGN/V1 response in the next layer HVA due to its amplifying effect on the neuronal activity. The effect can be seen by comparing the HVA response with attention to a condition where attention is disabled (Fig. 5.8 left versus right). HVA is simulated by the microcircuit model of attention that models this effect through the amplification mechanism. The mechanism relies on an amplification signal A that multiplicatively increases the response R of an neuron (Eq. 5.12). The signal A is zero if attention is disabled. Besides the amplification, the microcircuit model predicts further a weak response enhancement even without attention due to its divisive normalization mechanism (Eq. 5.12).

$$R = \frac{E \cdot (1 + A)}{\sigma + E \cdot (1 + A)} \tag{5.12}$$

Whereby R denotes the response in HVA, E the excitation from LGN/V1, and A the amplification through attention. The parameter $\sigma = 0.3$ controls the shape of the response function (Chap. 3).

The divisive normalization mechanism enhances low excitation values, resulting in an additional extension of the HVA response over time. This effect can be seen by comparing the excitation signal and the HVA response in the condition without attention (Fig. 5.8 right). Therefore, HVA maintains the transient signal weakly via the normalization mechanism without attention, and strongly via the combined influence of amplification and normalization with attention. Thus, we see the amplification mechanism as the primary source to maintain neuronal responses.

In summary, the maintaining of neuronal representations is based on the transient neuronal response in LGN/V1 and the extension of this signal in HVA via the amplification mechanism.

Set size effect due to spatial amplification

The set size effect results from the much earlier emergence of spatial attention in conditions with few items (e.g. C3) than in conditions with many items (e.g. C2). As outlined in Sec. 5.3.2, the processing in FEF converges in these cases differently fast to an activity

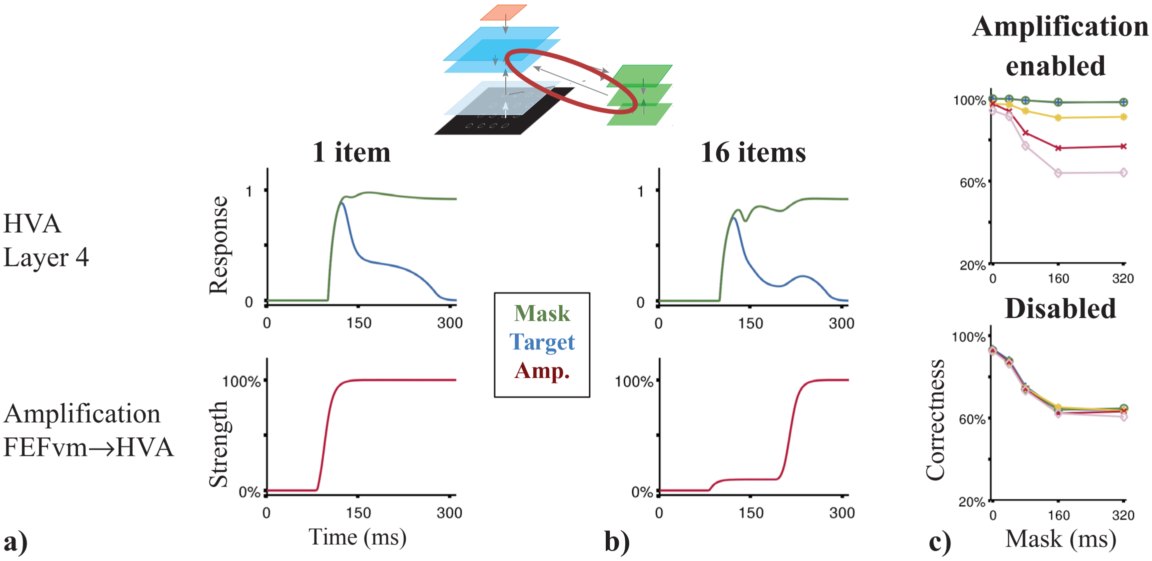


Figure 5.9.: The spatial amplification mechanism accounts for the set size effect. The mechanism involves FEFvm and HVA layer 4, indicated by the red circle in the model. The neuronal responses in HVA layer 4 are amplified at different time points in the conditions with 1 item **(a)** and with 16 items **(b)** by the signal from FEFvm. **c)** Disabling this mechanism removes the set size effect as all items show the same performance.

blob in FEFvm at the target location. This activity blob results in an amplification signal towards the target location (spatial amplification mechanism) and in a suppression of all other location (spatial suppression mechanism). Both signals are transported via connections from FEFvm to HVA layer 4.

Here, we are interested in the spatial amplification as this mechanism increases the neuronal response of the target. We consider here the two trailing mask conditions C2 and C3, thus neurons encoding the mask receive a constant strong excitation and neurons encoding the target a declining one. In a condition with one item (C3), the amplification signal appears very early and then rises strongly (Fig. 5.9a). This amplifies both target and mask excitation signals, leading to a high mask response and to the important increase of the target response. Contrary in the condition with 16 items, the amplification remains on a very low amount and increases only lately (Fig. 5.9b). Thus, the target response decreases swiftly due to missing amplification and received strong suppression. When the late amplification signal finally arrives, the target response is already very weak and the amplification has only a negligible influence. Thus, the target response remains low over the whole trial, causing the masking effect in OSM.

The set size effect is classically known from visual search where larger set sizes lead to higher reaction times (Wolfe, 1994). Hamker (2005a) explains this by a late emergence of spatial attention, which was indeed reported by Cohen et al. (2009). The late emergence of spatial attention is similarly observed in OSM, so we conclude that the set size effect in OSM is based on the same attentive dynamics as the set size effect in visual search.

To verify that the set size effect is really induced by the spatial amplification mechanism, we disabled it by cutting the connection FEFvm→HVA layer 4. The result shows an identical performance in all set sizes (Fig. 5.9c), proofing that the set size effect is indeed caused by this mechanism.

5.4. Discussion

In the following, we will relate the proposed model to existing models and theories of OSM, and discuss two currently open research questions: the necessity of reentrant processing, and the independence of the two factors mask duration and set size.

5.4.1. Relation to existing models and theories of OSM

We will discuss the computational model of object substitution (CMOS), the object updating hypothesis, and Põder's model with divided spatial attention. We examine the first two as they constitute the two major theories of OSM according to Goodhew et al. (2013), and the last one due to its relation to attention. The basic idea behind CMOS is, according to their authors, reentrant processing (Di Lollo et al., 2000), yet Põder (2013) and we found that reentrant processing is not correctly implemented in the model. Thus, we will treat the idea and the implementation of CMOS separately, and afterwards relate our model to the implementation of CMOS.

CMOS and the idea of reentrant processing

The computational model of object substitution (CMOS) from Di Lollo et al. (2000) was the first model explaining OSM and promotes prominently the theory of reentrant processing. The idea is that a feedback signal is projected back from a high-level area to a low-level area, so the high-level stimulus representation can reenter the low-level area.

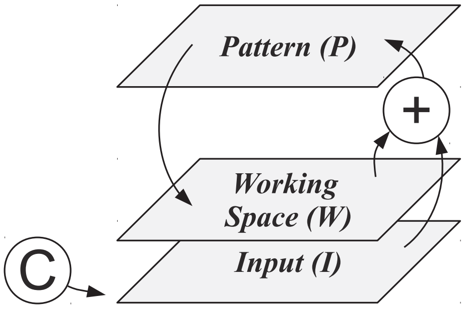


Figure 5.10: Computational model of object substitution (CMOS), reprinted from Di Lollo et al. (2000). It consists of an input layer (I), a low-level area (working space W), and a high-level area (pattern memory P). The two latter ones are recurrently connected, implementing reentrant processing.

The feedback forms together with feedforward signals the reentrant processing. Di Lollo et al. (2000) proposed reentrant processing for solving two different ambiguities. Firstly, the high-level area has a low spatial resolution, so its content has to be projected back to resolve spatial ambiguities. Secondly, multiple possible representations of the input in the high-level area have to be resolved in favor of the best one.

The low-level area is represented in CMOS (Fig. 5.10) by two layers, an input layer I and a low-level working space W. Di Lollo et al. (2000) speculated that the layer I might be represented in the cortex by area V1 layer 4, and W by V1 layer 2/3. The high-level area is represented by a high-level pattern memory P, which might be represented by area V3 or V4. The involvement of a low- and of a high-level area is supported by Di Lollo et al.'s findings of a low-level and of a high-level masking process. Each area in the model consists of spatially local modules, similar to the idea of cortical hypercolumns (Hubel et al., 1978), which process the different items in the display independently.

CMOS operates in the following way: the stimulus is encoded by I, whose content is initially fed to P. The content of P is projected back to W with a delay, forming a reentrant signal. W represents the same content as P, but in a high spatial resolution. The input (I) is then compared with W to resolve the previously outlined ambiguities, and the result is sent to P. The loop between P and W implements the proposed reentrant processing, and maintains neuronal representations too.

CMOS explains the impaired recognition accuracy in OSM by a mismatch between the reentrant signal from the high-level area P in W and the low-level representation in I. Initially, the high-level area P holds in all conditions a strong representation of mask and target, which originate from the simultaneously presentation of both stimuli at the beginning of the experiment. In the trailing mask condition C2, mask and target representations in P are projected to W, however the target is already gone in the input layer I when these signals arrive. The resulting mismatch between W and I is resolved in favor of the mask, strengthen the mask and weaken the target representation. The latter leads to the low target recognition accuracy in OSM. Contrary in the common offset condition C1, the input I contains a weak, decaying representation of both mask and target. These representations match with the mask and target signals from P. Therefore, the target representation is preserved, leading to the observed high recognition accuracy in C1.

Implementation and attentional gating view of CMOS

Regrettably, CMOS implements in our and Põder (2013)'s opinion the previously outlined idea of reentrant processing much too simply. Before we discuss this flaw, we will give the complete description of the model. The model implements the three layers (input layer I, working space W, and pattern memory P) in a separate module at each item location (Eq. 5.13, 5.14). Each layer contains three separated signals for target (j=1), mask (j=2), and internal noise (j=3). Time is simulated via discrete iterations (t), each representing a certain time span in the experiment.

$$P_j(t) \quad = \quad W_j(t-1) + \lambda \cdot I_j(t-1) \tag{5.13}$$

$$W_j(t) \quad = \quad \frac{P_j(t-1)}{\sqrt{\left(\sum_{j'} P_{j'}^2(t-1)\right)}} \tag{5.14}$$

Whereby the index j denotes the signal slot and the index t the iteration.

Spatial attention is assumed to be necessary for a correct recognition, thus the model iterates until attention is deployed and then reads out the signals in P. This read-out time point is denoted as t_c. The model assumes a late deploy of spatial attention in condition C2 and an early deploy in C3. However, the time point (t_c) is fitted directly by the data with a linear function of the set size (Eq. 5.15), and thus ignores known non-linear effects of the set size (sample-size law, Lindsay et al. (1968); Põder (2013)).

$$t_c \quad = \quad S \cdot n \tag{5.15}$$

Whereby S denotes a free search-rate parameter and n the set size.

The probability of recognizing a target correctly depends on the target signal compared to all other signals in P at the read-out time point t_c (Eq. 5.16).

$$Prob_{correct} = \frac{1}{K} \cdot \frac{P^2_{target}(t_c)}{\sum_{j'} P^2_{j'}(t_c)} \tag{5.16}$$

Whereby K is the asymptotic value for which $Prob_{correct} = 1$. This free parameter was fitted to the data, resulting in $K = 0.475$.

Unfortunately, the reentrant processing described by Di Lollo et al. (2000) is not implemented in CMOS due to too much simplification. In the implementation, the comparison of the reentrant signal with the low-level representation is missing, and thus the key component of reentrant processing is not specified in the model. According to the model equations, W is a delayed and normalized copy of P (Eq. 5.14), and P is a sum of W and I (Eq. 5.13). None of them implements a comparison operation.

This flaw was nicely elaborated by Põder (2013). He reformulated Eq. 5.13 and 5.14 to:

$$P_j(t) = \frac{P_j(t-2)}{\sqrt{\left(\sum_{j'} P^2_{j'}(t-2)\right)}} + \lambda \cdot I_j(t-1) \tag{5.17}$$

From this, it is possible to see that P implements a trace which follows the input I with some inertia, but does not implement reentrant processing. The trace is used in the model to maintain a representation of mask and target stimulus after their offset. Hence, Põder concluded that the model implements solely the following components: maintaining of representations, normalization, and attentional delay. Despite the missing reentrant processing, the model accounts well for the OSM data. Thus, Põder (2013) concluded that only these components are necessary to explain OSM, but not reentrant processing.

Therefore, the implementation of CMOS simulates OSM not with the help of reentrant processing: In the common offset condition C1, the target and mask representations in P decay similarly. At the arrival of spatial attention, the target representation is still strong enough to ensure a correct behavioral response (Eq. 5.16). Contrary in a trailing mask condition C2, the mask remains and only the target decays in P. Due to the normalization among all representations in P (Eq. 5.17), the target signal is diminished by the strong mask signal which leads to an incorrect decision.

Furthermore, Põder (2013) saw CMOS as a model of attentional gating (Reeves and Sperling, 1986; Sperling and Weichselgartner, 1995; Smith and Ratcliff, 2009) as shown in Fig. 5.11. In such a model, a stimulus must be attended in order to perceive it, hence attention

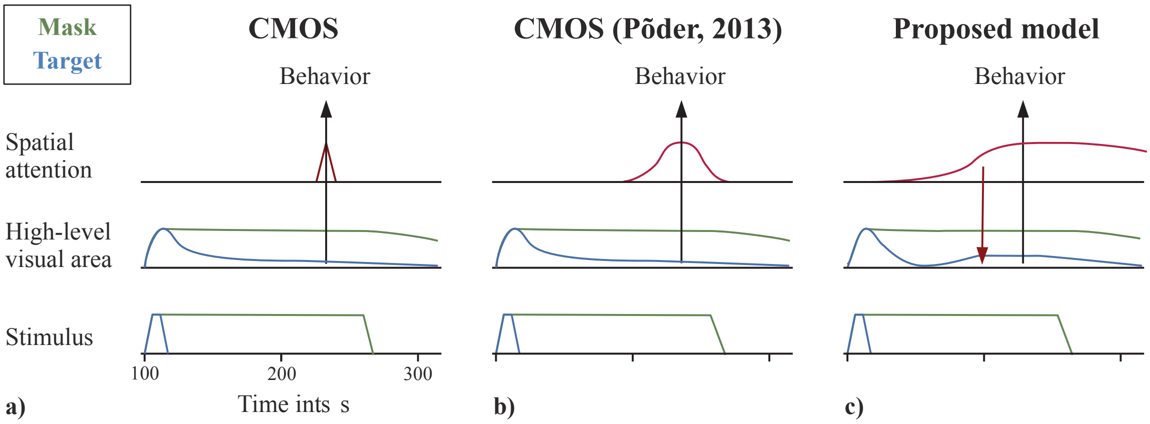


Figure 5.11.: Comparison of the processing within the different models. For each model, the stimulus, the highest visual area which holds the stimulus representation, and spatial attention is illustrated. The stimulus is set on at 0 ms, but is displayed aligned with the onset in the highest area. The highest area is in CMOS the pattern memory P, and in the proposed model HVA layer 2/3. For clarity, all curves are schematic illustrations and do not show actual model results. The setup uses a trailing mask of 160 ms and 16 items, a condition which would result in an impaired accuracy. **a)** Implementation of CMOS by Di Lollo et al. (2000). **b)** Attentional gating view of CMOS by Põder (2013). **c)** Our proposed model. The models differ at most in the role of spatial attention. Di Lollo et al. see attention as a mechanism to read out the neuronal target response at a singular time point (a). Contrary, Põder sees attention as a gate, gating continuously neuronal information to behavior (b). In our model, attention amplifies the target response continuously over time (c).

gates the stimulus information to higher cortices to evoke a behavioral response. In CMOS, the target representation is read out to determine the behavioral response when attention is deployed (Eq. 5.16, Fig. 5.11a). Põder viewed this mechanism as a gate (Fig. 5.11b) and concluded that CMOS is an attentional gating model. In our theory, attention also facilitates perception as it amplifies the neuronal representation (Fig. 5.11c). The difference is that we do not see attention as a gate, instead we stress the physiological amplification effect at the neuronal response. Thus, the difference is that Põder's view needs attention for a correct decision, whereas our view requires a high firing rate, which can result from attention, but also from different causes like a strong input or a low amount of suppression.

Relation of CMOS to our proposed model

Our attention model has several advantages compared to CMOS: Firstly, our model simulates the visual cortex much more sophisticated and thus can use real images as input. Secondly, CMOS explains solely OSM, whereas our model explains attention paradigms at neuronal level like biased competition, guided visual search, as well as OSM. Thirdly, our model is, contrary to CMOS, underpinned with a large corpus of neuroanatomical data like realistic receptive field sizes, synaptic transmission delays, cortical connectivity and areas, and top-down influences via attention. As our model is much more solidly-grounded than CMOS, we will now use it to verify the components of CMOS. We will consider here the model implementation of CMOS as the OSM data was fitted by it. This implementation contains the components: normalization, attentional delay, and maintaining of neuronal representations.

In both models (Fig. 5.11a, c), the masking effect occurs in condition C2 because the neuronal representation of the mask suppresses the representation of the target (Sec. 5.3.3). However, the implementation of the suppression is quite different. CMOS uses a normalization operation in W (Eq. 5.14), whereas we see suppression as a neuronal mechanism of attention (Sec. 5.3.3). These operations might be equivalent as CMOS uses a much higher level of abstraction than our model.

Both models propose a delayed onset of spatial attention in the trailing mask condition C2 and an early onset in C3. Besides this, the models cope with spatial attention very differently. In CMOS, attention denotes the time point to read out the neuronal target response (Fig. 5.11a). The time point is determined by a linear function dependent on the set size (Eq. 5.15). Further attentive dynamics are not simulated. Hence, CMOS does not specify the source of spatial attention, thus we see it is external signal. Contrary in our model, spatial attention emerges from a recurrent loop between HVA and FEF (Sec. 5.3.3). The time point of its onset depends on the spreading of the FEF activity, and thus it is delayed in C2 as the 16 items cause a broad activity. In our model, spatial attention has not the role to determine the read-out time point as in CMOS, instead it amplifies the neuronal target response and influences in this way the behavior (Fig. 5.11a). Therefore, our model includes a rich set of attentive dynamics, whereas Di Lollo et al. (2000) set spatial attention externally.

A big difference between the models lays in the way how the representation of a stimulus is maintained after its offset. CMOS implements this via reentrant processing in the sense that the stimulus representation is propagated back and forth with a delay and so it decays only slowly. We propose a much simpler solution: Spatiotemporal receptive fields in LGN/V1 evoke a swiftly fading response that is maintained locally in HVA by the amplification mechanism (Sec. 5.3.3).

To conclude, our model supports the three components of CMOS only partially. The first, the normalization, seems to be compatible with our model. The second, the delayed onset of spatial attention in C2, is part of our model, but much simpler implemented: CMOS sets attention externally, whereas we simulate a rich set of attentive dynamics. The third, the maintaining of representations, cannot be confirmed. In CMOS, representations are maintained via reentrant processing, which is not required in our model. Hence, we reason that reentrant processing is not necessary to maintain neuronal representations in OSM.

Non-computational theories of OSM

We discuss now models and theories without a computational implementation. Such theories typically do not simulate stimulus representation or attention signals over time, making their concrete realization in the cortex rather vague. This makes it impossible to profoundly verify them by neuronal mechanisms or by our model. Thus, we discuss them very briefly and encourage the original authors to further expand their theories.

The model of Põder (2013) is interesting due to its relation to spatial attention. The model extends CMOS by a divided stage of attention, in which spatial attention is distributed across all locations. The divided attention stage describes the time span in OSM before attention is deployed to the target location. The time span afterwards is denoted by Põder as selective attention stage. Classically, CMOS presumes that information from the stimulus is only read out in the selective attention stage. However, Põder assumes that information is extracted in the divided stage too. Thus, the behavioral performance depends in his model on the information extracted from both stages. In our model, the performance depends on the integrated target response over the full experiment duration, which includes also information before spatial attention is deployed. Therefore, we conclude that Põder's approach is compatible with our model.

The object updating hypothesis is the second most popular OSM theory besides feedforward or reentrant models (Goodhew et al., 2013). It was firstly formulated by Lleras and Moore (2003) and by Moore and Lleras (2005), and later generally described by Pilling and Gellatly (2010). The theory explains OSM by objects competing for a representation in mid- or high-level visual cortices (Pilling and Gellatly, 2010). The masking effect occurs if target and mask are encoded as a single representation. The theory assumes that objects are constantly updated, thus the target is updated and by this becomes the mask. When this can be prevented, two representations emerge and OSM does not occur. This is the case with a mask moving outwards, sliding past the target, jiggling independently of the target, or having a different color than the target (Lleras and Moore, 2003). The idea, that masking occurs if the target representation is incorrectly updated, originates back to Di Lollo et al. (2000). They stated that OSM occurs because an initial display (target and mask) on the screen is replaced by a different configuration (mask alone), leading to a replacement of the initial percept.

The object updating hypothesis is not included in our model, but seems not contradictory to it. We cannot relate our model to the updating process as the theory does not specify its neural mechanisms, but we could evaluate the resulting object representations in our model and in the theory. Object representations are currently not part of our model. For our evaluation, we would assume that they are stored in some form of visual short term memory (VSTM, Sperling (1960); Emrich et al. (2011); Todd and Marois (2004); Todd et al. (2005)), and imagine that we have add a VSTM above HVA. We would further assume that the probability of storing a representation in VSTM is equal to its neuronal response in HVA layer 2/3. With the latter assumption, the VSTM would contain mask and target representation in the common offset condition C1 as both responses are strong in HVA (Sec. 5.3.2), and solely the mask in the trailing mask condition C2 as only this response is strong (Sec. 5.3.2). The object updating hypothesis predicts the same outcome: target and mask representation in C1, and solely the mask in C2. Therefore, the hypothesis and our model seem to be compatible.

5.4.2. Discussion of open questions in OSM

Necessity of reentrant processing

Di Lollo et al. (2000) prominently promote the idea of reentrant processing in OSM. Its function is to verify a high-level representation with a low-level representation to resolve ambiguities. Such verification is firstly necessary to resolve ambiguities if multiple high-level representations are possible like in the case of illusions. Secondly, high-level areas have a low spatial resolution, thus a mapping to the precise location is required. However in their model, the function of reentrant processing is only to maintain the neuronal target representation after stimulus offset (last section).

Other researchers swiftly argued against the new reentrant theory and in favor of current masking theories. Francis and Hermens (2002) showed that OSM can be simulated by known models of metacontrast masking (Bridgeman, 1978; Francis, 2000; Weisstein, 1968), and concluded that a novel reentrant theory is not necessary. However, Di Lollo et al. (2002) disagreed and stated that Francis and Hermens (2002) modeled spatial attention in an inappropriate way. In Francis and Hermens (2002), attention does not affect the target, but decreases the mask intensity and so weaken the masking effect for low set sizes. Di Lollo et al. (2002) argue that this approach assumes spatial attention to the mask, but not to the target. This assumption is contradictory to the coarse resolution of spatial attention, which lies in the range of degrees (He et al., 1997; LaBerge, 1990) and not in the range of arc-minutes like the separation of target and mask ($6'$ in experiment 1 of Di Lollo et al. (2000), $0'$ - $36'$ in experiment 3). Thus, the mask could not be attended independently from the target.

Põder (2013) reinvestigated the work of Di Lollo et al. (2000) and stated that the paper indeed presents a reentrant theory, but that this theory is not implemented in the CMOS model (Sec. 5.4.1). Instead, the model implements only a trace following the input, which is used to maintain stimuli representations. As the model can perfectly account for the OSM data, Põder concluded that OSM can be explained without reentrant processing.

In the study of Jannati et al. (2013), Di Lollo disagrees at Põder (2013)'s explanation of OSM. However, his main arguments are against the divided-attention-stage model of Põder (Sec. 4.1.4) and not against his view of CMOS. Põder's model predicts a decreasing recognition accuracy with increasing mask duration, and that the accuracy is unaffected by an interstimulus interval (ISI) between the simultaneously presented mask plus target and the mask alone. Jannati et al. (2013) tested the latter prediction and found it not to be true. They used the four-dot experiment of Di Lollo et al. (2000), added a variable interval between target and mask alone, and decreased the mask duration to 10 ms. They found that the accuracy is primary impaired at ISIs around 80 ms and recovers for greater ISIs. The value of 80 ms indicates that the mask suppresses the target primary at 80 ms after target offset. They interpreted this as the time of reentrant signal, thus stated that the study supports Di Lollo et al.'s theory. However, our model could account for the data without this interpretation and without supporting reentrant processing. Before 80 ms, the brief mask would have a weak effect as this condition is similar to the short mask durations in the original experiment, and our model can account for it very well (Sec. 5.3.1). For longer ISIs than 80 ms, the spatial amplification signal from FEF reaches HVA before the mask is set on again. This signal amplifies the target response and so cancels out the later suppression from the mask. This results in a strong target response and high accuracy. In our simulations, the spatial amplification signal starts to increase at 220 ms in HVA (condition C2, Fig. 5.5). A stimulus signal reaches HVA 100 ms after its onset, so the mask signal will arrive before the amplification for all ISIs less than 120 ms. Hence, the accuracy should begin to recover for all ISIs greater than 120 ms, which is indeed observable in the data.

Põder argued against Jannati et al. (2013) in his next publication (Põder, 2014), defending his view of CMOS as attentional gating. However, we think Põder and Di Lollo misunderstood each other's arguments. Põder claimed that Di Lollo is against viewing CMOS as attentional gating, but Di Lollo defended primary his view of CMOS as reentrant model, which is unrelated to attentional gating. On the other hand, Di Lollo showed in Jannati et al. (2013) solely that Põder's model cannot account for his OSM data, but concluded from the incorrect model that Põder's claims about reentrant processing are wrong too. Yet, Põder's model is completely unrelated to reentrant processing.

In summary, two evidences indicate the non-involvement of reentrant processing in OSM, even though none of them rule it out: The CMOS implementation uses reentrant processing solely to maintain neuronal representations. Regarding this point, we can conclude from our model that reentrant processing is not necessary (Sec. 5.4.1). The remaining aspect of reentrant processing, i.e. the resolving of ambiguities, was neither implemented in CMOS nor in our model, so its role in OSM remains unclear. However, our model does not require reentrant processing to explain OSM as attention phenomenon. Regarding the discussion in the literature about reentrant processing, the debate is not finished yet and there seems to be an ongoing misunderstanding between Di Lollo and Põder. It is currently hard to vote for one side as the outlined arguments are supported by evidences, but are not included in CMOS or Põder's model, and so could not be tested. Thus, we hope for a further development of these models.

Independence of mask duration and set size

Di Lollo et al. (2000) identified mask duration and set size as the two major factors affecting the recognition accuracy. However, it remains unclear if these two factors are independent or not. Di Lollo et al. (2000) stated that these factors have a 'joint' influ-

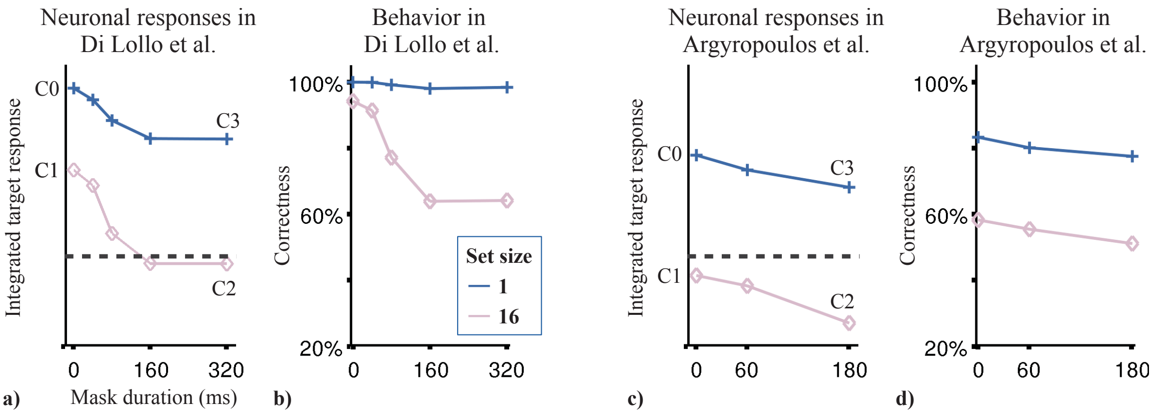


Figure 5.12.: Di Lollo et al. (2000) reported a dependency of the factors mask duration and set size (b), whereas Argyropoulos et al. (2013) found an independence (d). We found that the factors are indeed independent at neuronal level in OSM (a, c), and are observed only dependent in the behavioral data of Di Lollo et al. (2000) due to a ceiling of the curves at small set sizes (b). **a)** At neuronal level, the evidence for recognizing the target correctly is represented by the integrated neuronal response of the target gap neuron. For clarity, only the extreme conditions with a set size of 1 and 16 items are shown. If the evidence exceeds a perceptual threshold, the model recognizes the target correctly. Otherwise, a random decision is taken. The threshold is chosen trial-wise from a Gaussian distribution, whose mean (fine-dotted gray line) is identical in both experiments (a, c). **b)** Resulting behavior of the model in Di Lollo et al. (2000). **c, d)** The integrated neuronal response (c) and the resulting behavior (d) in Argyropoulos et al. (2013).

ence, and concluded from the mean-accuracy results (Fig. 5.12b) that they are dependent. Unfortunately, they concluded this dependency for an experiment with a full-ring mask (denoted as experiment 1 in Di Lollo et al. (2000)), and proposed that it is also valid for the here considered four-dot mask (experiment 3). As another issue, they did not run an analysis of variance test (ANOVA) for the here considered experiment 3, only for the later experiments 4 - 6. The later experiments use a target that is easier recognizable, and their ANOVA results indeed show the significance of the dependence.

Argyropoulos et al. (2013) interpreted these evidences so that both factors are dependent in all experiments of Di Lollo et al. (2000). Argyropoulos et al. (2013) ran a new study (Sec. 5.3.1) and found to their surprise the independence of the two factors (Fig. 5.12d). They also verified the significance of their results via ANOVA tests. Their study avoids floor and ceiling effects at the recognition accuracy curves. They concluded that both factors are normally independent and appeared only dependent in Di Lollo et al. (2000) due to a ceiling of the accuracy curves for small set sizes.

We conclude from our model that the two factors affect the recognition accuracy independently via two different neuronal attention mechanisms: The factor mask duration influences the accuracy via the suppression mechanism (Sec. 5.3.3), whereas the factor set size affects the recognition accuracy via the spatial amplification mechanism (Sec. 5.3.3). Therefore, we would normally expect independence as shown by Argyropoulos et al. (2013). To inspect why Di Lollo et al. (2000) shows contrary a dependency, we examine the neuronal target response underlying the behavior in both studies (Fig. 5.12). In our model, the integrated neuronal response of the target gap neuron determines the behavior as it represents the accumulated, neuronal evidence for recognizing the target (Sec. 5.2.2). If the evidence exceeds a perceptual threshold, the model recognizes the target correctly, otherwise it takes a random decision. The threshold is chosen trial-wise from a Gaussian distribution, whose mean is identical for both experiments. We found that the integrated neuronal target responses show an independence in both experiments (Fig. 5.12a, c). Yet, the integrated neuronal responses in Di Lollo et al. (2000) are in many conditions, especially for small set sizes, over the perceptual threshold (fine-dotted line), leading to a perfect behavior in these cases. Due to this ceiling effect, the factors are observed dependently at behavioral level (Fig. 5.12b), despite they are independent at neuronal level. Contrary in Argyropoulos et al. (2013), the integrated neuronal responses are lower as their experiment is harder (Sec. 5.3.1). Hence, the integrated responses are not always over the varying perceptual threshold, avoiding the ceiling effect. As the ceiling is avoided, the behavioral curves show now the independence of mask duration and set size (Fig. 5.12d), similar as in the underlying neuronal responses. Therefore, we can confirm Argyropoulos et al. (2013)'s explanation that the dependence in Di Lollo et al.'s original data is due to the ceiling effect.

To further verify that finding, we model a variation of Di Lollo et al. (2000)'s experiment that avoids the ceiling effect. For this, we increase the neuronal threshold for a correct decision in the model (R_x in Eq. 5.1) to make the task more difficult. Experimentally, this could be realized by a change in the experimental instructions like that subjects should only reported a decision if they are very confidently. We found that this variation of Di Lollo et al.'s experiment shows independence, confirming again Argyropoulos et al. (2013)'s explanation.

However, both factors could theoretically also be dependent as their underlying mechanisms affect the same cells in HVA layer 4. We evaluated each extreme conditions (C0, C1, C2, C3, Fig. 5.12) at neuronal level in the experiments of Di Lollo et al. (2000) and

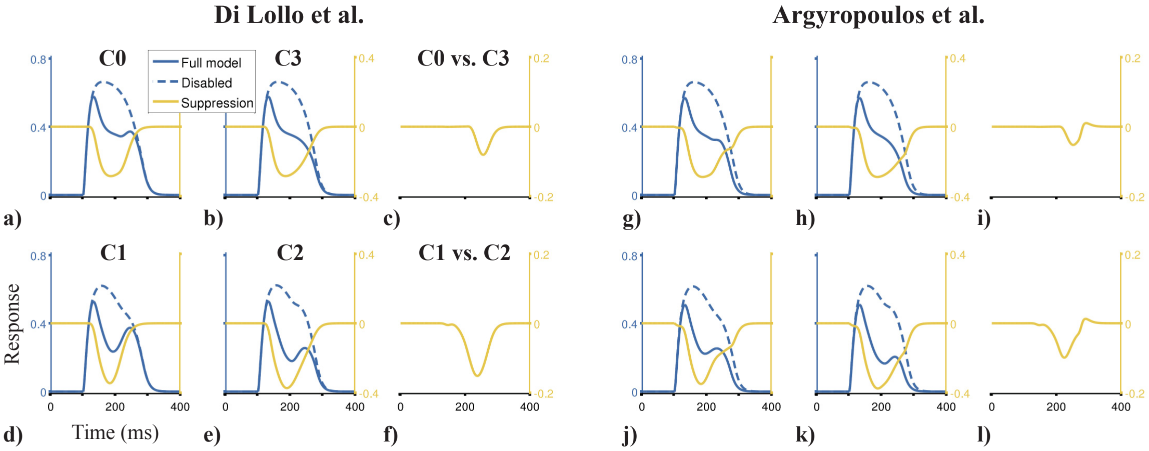


Figure 5.13.: Net effect of the suppression on the neuronal response. **a,b,d,e)** For each condition C0 - C3 in Di Lollo et al. (2000) (Fig. 5.12), the suppression effect is plotted as the difference between the neuronal response of the full system and the response with the suppression mechanism disabled. For disabling, the suppressive connection from HVA layer 2/3 to layer 4 was cut. **c,f)** The mask duration effect is shown by the difference of the suppression which is greater in the trailing mask conditions (C2 and C3). **g-l)** The suppression effect in Argyropoulos et al. (2013), similarly plotted as in a-f).

Argyropoulos et al. (2013), and found that especially C1 (16 items, 0 ms mask) differs slightly, but notable between both experiments: The integrated neuronal response in C1 is relatively lower in Argyropoulos et al. (2013) than in Di Lollo et al. (2000) (Fig. 5.12a vs. c). Thus, the factor mask duration decreases lesser the behavioral performance in Argyropoulos et al. (2013) (C1 vs. C2). The lower C1 response results from a greater effect of the suppression (Fig. 5.13d vs. j), especially late around 250 ms. The suppression signal from HVA layer 2/3 to 4 is identical in both cases, but its influence on the response in layer 4 differs because of the divisive normalization function (Chap. 3). In this function, the influence of the suppression depends non-linearly on the excitation. In Argyropoulos et al. (2013), the excitation is weaker than in Di Lollo et al. (2000) due to the smaller gap in the target ring (Sec. 5.3.1). The changed excitation increases the effect of the suppression in all conditions as observed by the lower rates around 250 ms (Fig. 5.13g - k vs. a - f), but most prominently in C1. Thus, the suppression increases less in Argyropoulos et al. (2013) (Fig. 5.13i, l) than in Di Lollo et al. (2000) (Fig. 5.13c, f) with longer mask durations. Hence, there exist weak a interaction between mask duration and set size, but it decreases the mask duration influence in Argyropoulos et al. (2013) and so even enhances the independence of both factors in this study.

In summary, the factors mask duration and set size are independent. We found two aspects influencing their independence. The primary one is the ceiling effect as proposed by Argyropoulos et al. (2013), and the other one is the neuronal interaction of excitation, feature-based suppression, and spatial amplification in HVA.

5.5. Conclusion and outlook

5.5.1. Conclusion

We demonstrated that our previously-developed model of visual attention (Chap. 4) is able to account for object substitution masking (OSM): We showed that the model fits the typical behavioral data sets of OSM (Sec. 5.3.1), explained the neuronal dynamics in OSM (Sec. 5.3.2), and linked these explanations to the neuronal mechanisms underlying visual attention (Sec. 5.3.3). Therefore, we conclude that the different phenomena of visual attention and OSM rely on the same underlying, neuronal mechanisms. Hence, OSM can be explained by well-known theories of visual attention, indicating that there is no need for a separate OSM theory.

We found out that especially three neuronal attention mechanisms are important for OSM (Sec. 5.3.3): 1) The impairment of the behavioral performance is explained by a feature-based suppression mechanism between mask and target. The same mechanism explains the reduction of neuronal activity in many attention experiments, for example in the well-known biased competition paradigm (Chap. 3). 2) The neuronal target response has to be maintained after stimulus offset in OSM. We explain this by the amplification of neuronal responses, which is caused by attention (Chap. 3). Here, the mechanism amplifies the rapidly decreasing neuronal response induced by spatiotemporal receptive fields in a low-level, presynaptic area, and so effectively extends the response. 3) The set size effect is explained by the temporal dynamics of spatial attention. Spatial attention enhances the neuronal responses by a spatial amplification mechanism, and emerges naturally in our model by the networks dynamics between the higher visual area (HVA) and the frontal eye field (FEF). With a few items, attention emerges early at the target location, thus the neuronal target response is amplified from early on, leading to a highly correct behavioral performance. In the condition with many items, attention is weak and emerges late, thus the amplification mechanism's influence on the response is negligible, leading to a lower performance. These temporal dynamics of spatial attention stem from attention models explaining visual search, where the late onset of attention accounts for the longer search times with many distractors (Hamker, 2005a). By explaining OSM via these attention mechanisms, we are able to transfer our understanding of visual attention to visual masking, and hope we further bridge the gap between these two fields.

Furthermore, we relate our model to major existing theories of OSM, the computational model of object substitution masking (CMOS) and the object updating hypothesis (Sec. 5.4.1). In relation to them, our neuro-computational model has the advantages that it is strongly rooted by neurophysiological and neuroanatomical findings, and that it explains a large set of attention paradigms in addition to OSM. These facts create a much stronger foundation of its inherent mechanisms than in current OSM models, thus it represents probably the best neurologically-grounded model for OSM.

As our model is much more solidly-grounded, we used it to verify the existing theories of OSM: CMOS and the object updating hypothesis. CMOS proposes the idea of reentrant processing, but simplified out most of this idea in the implementation. This implementation explains the masking effect similar to our suppression mechanism, but is simpler in all other aspects. For example, spatial attention emerges intrinsically in our model from the network dynamics, whereas it is a simple external signal in CMOS. Some aspects of CMOS can also not be confirmed, for example reentrant processing is not required to maintain neuronal representations. The object updating hypothesis explains OSM at the abstract level of memorized object representations, whereas we explain it at neuronal level. We worked out that both explanations are probably compatible.

Finally, we address open research questions with our modeling approach (Sec. 5.4.2): Firstly, we gather evidences that reentrant processing between a low- and a high-level visual area might not be necessary to explain OSM. Secondly, we conclude that the factors mask duration and set size influence independently the behavioral performance, as they utilize separate attention mechanisms. This independence was already proposed by Argyropoulos et al. (2013), so we endorse their theory. The data of Argyropoulos et al. (2013) show this independence, whereas Di Lollo et al. (2000) show a dependency due to a ceiling of the accuracies (Argyropoulos et al., 2013). Our model replicates both data sets.

5.5.2. Outlook

The model opens many future research directions as it is the first neural system-level model of OSM, creating the opportunity to model neuronal activity as well as EEG data. There exists already EEG data sets obtained in OSM (Crouzet et al., 2014; Pagano, 2012; Prime et al., 2011; Woodman and Luck, 2003; Woodman, 2010), thus we could now model these data sets to better investigate the temporal dynamics of OSM.

Particular interesting of the EEG data is the N2PC component as it is associated with spatial attention and processing (Eimer, 1996; Luck and Hillyard, 1994). The N2PC is an EEG wave observed contralateral to an attended location, typically 200ms after stimulus onset (Luck and Hillyard, 1994), and is a negative, posterior recorded EEG component. Classically, the N2PC is supposed to be an indicator for spatial attention (Eimer, 1996; Luck and Hillyard, 1994). However, a recent study reported discrepancies, Kiss et al. (2008) observed that the N2PC is not modulated by cueing a location, whereas spatial attention is typically rapidly deployed to a cued location. In OSM, the temporal onset of the N2PC is not affected by set size (Prime et al. (2011); Woodman and Luck (2003), personal communication with S. Pagano about the study of Mazza and Pagano (2014)), whereas the temporal onset of spatial attention is affected (Sec. 5.3.3). The onset of the N2PC is also not affected in forward or backward masking paradigms (Robitaille and Jolicoeur, 2006). These discrepancies raise the research questions about the neuronal correlate of the N2PC, and about the precise connection between N2PC and spatial attention.

Future work could also include the modeling of other OSM variations. Especially interesting are variation in which two (Crouzet et al., 2014; Prime et al., 2011; Woodman and Luck, 2003; Woodman, 2010; Pagano, 2012) or multiple four-dot masks are used (Chakravarthi and Cavanagh, 2009; Tata and Giaschi, 2004). In these setups, the target is indicated by a conjunction of the 4-dot mask and an additional object category. This would imply a huge processing difference in our model. The distinct object category would allow to search the target directly via this feature, hence these OSM variants would be a guided visual search for the target. Contrary in the here used setup with a single four-dot mask, the target is only defined via the mask (Sec. 5.3.2). Thus, this OSM variant is initially a guided visual search task for the mask and not for the target, and only afterwards, the target is attended. These task differences imply contrary temporal dynamics, which might be very interesting to investigate. It might also resolve the discrepancy in the behavioral results of OSM. For example, Crouzet et al. (2014) reported similar performance curves for a two-masks OSM task and backward masking, whereas Enns (2004) reported different curves for OSM with a single-mask and backward masking.

6. General conclusion

6.1. General discussion

In this discussion, we will focus on cross-phenomena aspects as the modeling of an individual phenomenon is already discussed in each of the previous chapters (Chap. 3.4, 4.5, 5.4 respectively). We will address the following points: Firstly, we will compare how the neuronal mechanisms of attention operate in the two behavioral phenomena. Following this, we will look at the roles of the mechanisms in neurophysiology and behavior to bridge the gap between both fields. Finally, we will discuss the boundaries of our work.

6.1.1. Role of the neuronal mechanisms in behavior

We observe differing roles of the neuronal attention mechanisms in the two behavioral phenomena, i.e. in object localization and in object substitution masking (OSM). We reason that these findings stem from the different task classes of the two phenomena: Object localization is a typical localization task, while OSM is a recognition task. To recapitulate, we have defined these task classes as follows (Chap. 1.1): In a localization task, the type, i.e. the feature of the target stimulus is given, and its location has to be determined. In a recognition task, the location of the target stimulus is given, and its type (i.e. feature) has to be determined. OSM is a recognition task as the target location is given by the four-dot pattern and the type of the target ring has to be determined.

We found out that the model engages the amplification and suppression mechanisms diametrically in the two task classes to meet their different demands. Regarding amplification, we observed the following: In the object localization task, the feature-based amplification is essential as it represents the given target feature in the visual system (Chap. 4.3.1). In contrast in the recognition task (OSM), the spatial amplification is crucial as it represents the given target location. We deduce this from the fact that the recognition performance in OSM is only impaired when spatial amplification is not available (Chap. 5.3.2).

Regarding suppression, we found that the mechanism of feature-based suppression has a completely opposing role in each of the two tasks. In object localization, it does not affect the task performance as disabling the mechanism does not impair the localization accuracy (Chap. 4.3.2). On the contrary, in OSM, it affects the performance as it accounts for the recognition impairment by masking (Chap. 5.3.3). Hence, we conclude that the mechanism must facilitate recognition tasks. The mechanism implements a competition between features, thus it selects the strongest feature from several alternatives. Therefore in recognition tasks, the suppression determines the best feature on a given location. Another study reported similarly the importance of the feature-based suppression for recognition tasks: Kermani Kolankeh et al. (2015) showed that competition improves classification performance of single handwritten digits under occlusion. Their task belongs to the recognition class as the location is given by the uniqueness of the stimulus and the target feature has to be determined. Their competition operates among the feature dimension, and in doing so realizes the same competitive effect as our feature-based suppression mechanism. Regarding the spatial suppression, the mechanism was introduced in the model for the object localization task (Chap. 4.2.3), thus we see it as crucial for localization tasks. The mechanism plays no role in the explanation of OSM (Chap. 5.3.3), thus we view it as irrelevant for this task. Hence, the mechanism seems to have no influence in recognition tasks. These findings indicate that also the spatial suppression mechanism has a diametrical role in the two task classes.

In our work, we model the primate brain's attentional processing, thus these observations shed light on how this processing may operate in different task classes. As attention is, according to our theory, a holistic and cognitive control process (Chap. 4.2.3), our insights predict how this control process may realize different tasks.

Put together, the model uses opposing spatial and feature-based variants of the mechanisms in both task classes. In recognition tasks, spatial amplification represents the target in the visual system via a given location, while feature-based suppression realizes a competition between features to determine the best one. In contrast in localization tasks, feature-based amplification represents the target via a given feature, while spatial suppression mediates a competition between locations to determine the best one. In both tasks, the amplification is balanced out by suppression, a finding similar to the idea of balancing out excitation with inhibition (Haider et al., 2006; Ozeki et al., 2004).

In summary, the amplification is always complemented with the opposing suppression mechanism, whereby amplification has the role to represent the target in the visual system

	1a) Feature-based amplification	1b) Spatial amplification	2) Normalization	3) Spatial pooling	4a) Feature-based suppression	4b) Spatial suppression	4c) Surround suppression
Biased competition with spatial attention		+		+	+		
Biased competition with feature-based attention	+				+		
Contrast gain		+	+				
Response gain		+	+				+
Scaling		+					
Sharpening	+				+		
Surround suppression		+					+
Object localization in guided visual search	+				-	+	
Objects substitution masking		+			+		

Table 6.1.: Overview of the neuronal mechanisms primarily involved in each attention experiment, denoted by "+". The symbol "-" denotes the case where we explicitly tested the non-involvement of a mechanism. The first row segment comprises neurophysiological experiments (Chap. 3), and the second segment behavioral experiments (Chap. 4, 5). The related explanation of a mechanism in a particular experiment can be found in Chap. 3.3, Chap. 4.4, and, Chap. 5.3 respectively.

and suppression to implement a competition between several behavioral alternatives to determine the best one. Due to this processing, our model is able to perform both task classes, predicting its applicability to a broad range of tasks.

6.1.2. Role of the mechanisms in physiology and behavior

We will now summarize the role of each neuronal model mechanism in the neurophysiological and behavioral phenomena (Tab. 6.1), and link in this way neurophysiology to behavior.

The amplification mechanism models, at the neurophysiological level, the response increase of neurons preferring the attended stimulus (Chap. 3.2.3). The attended stimulus is per definition task-relevant, thus the mechanism amplifies neurons encoding the target of the current task, and hence represents the task target in the visual system (last section 6.1.1). In object localization, it represents the target object via its features (feature-based amplification). In OSM, it represents the target via its location (spatial amplification) as masking occurs only if this mechanism is weakened. Aside from this, the amplification maintains the neuronal response in OSM (Chap. 5.3.3).

The divisive normalization mechanism regulates the neuronal response and thus has a primary neurophysiological role (Chap. 3.2.2). However, it might also facilitate the recognition of low contrasted stimuli as it increases weak neuronal signals, which result from low contrasts. For example, the author's preceding study of Antonelli et al. (2014) uses a normalization function in a much simpler form to localize low-contrast objects. In OSM, the signal enhancement by divisive normalization might explain why OSM is only marginally affected by stimulus contrast changes (Di Lollo et al., 2000; Luiga and Bachmann, 2008).

The spatial pooling mechanism is neurophysiologically used in biased competition where it transfers the spatial amplification in a part of the receptive field to the recorded neuron (Chap. 3.3.1). It has no notable effect in our behavioral experiments, but might have an effect in experiments evaluating spatial invariance as it is a typical mechanism to model it (HMAX, Riesenhuber and Poggio (1999); Visnet, Rolls (2012)).

The suppression mechanism inhibits the response of other neurons at the neurophysiological level (Chap. 3.2.4), for example in biased competition where it inhibits neurons encoding unattended stimuli (Chap. 3.3.1). At the behavioral level, unattended stimuli are per definition task-irrelevant, thus the mechanism suppresses neuronal activity evoked by such stimuli, like distractors or background clutter. We denote such task-irrelevant activity as neuronal noise, thus the mechanism removes as first function this noise and cleans in this way the neuronal code (Chap. 4.3.2). As second function, the suppression realizes, at the behavioral level, a competition between several behavioral alternatives to determine the best one (Sec. 6.1.1).

The suppression mechanism is modeled in three variations, i.e. suppression of neurons encoding different features (feature-based suppression), different locations in the frontal eye field (spatial suppression), or locations in the surround (surround suppression), thus we outline in the following additionally the role of each subvariant: Feature-based suppression is involved, at the neurophysiological level, in effects like biased competition or sharpening of neuronal tuning curves (Chap. 3.3.3). At the behavioral level, we reason that it is prominently relevant for recognition tasks like OSM, where it affects the behavioral performance, but is not relevant for localization tasks (last section 6.1.1). Spatial suppression is not part of the microcircuit model, thus it was not evaluated at the neurophysiological level. At the behavioral level, it plays an essential role in object localization as it removes activity at distractor locations, but according to our reasoning, has no role in recognition tasks like OSM (last section 6.1.1). The mechanism of surround suppression accounts neurophysiologically for two effects, surround suppression and response gain (Chap. 3.3.4). It plays a minor role in OSM, and it was disabled in object localization as a preliminary test showed that it was irrelevant for the task. However, it might play a more prominent role in tasks with multiple target stimuli, as suggested by surround suppression effects when storing multiple items in visual memory (Franconeri et al., 2013).

6.1.3. Boundaries of the conducted research

We will now discuss the boundaries of our work. However, as all aspects related to the modeling of a particular phenomenon have been already outlined in each of the previous chapters, we will discuss here only the few remaining, general points.

Firstly, our work does not cover the simulation of spiking neurons as we use mean-rate coded neurons (Chap. 3.2, 4.2, 5.2). Therefore, our model cannot simulate the temporal or short time-scale effects of visual attention (Chap. 2.3). We choose the level of mean-rate coded neurons as it allows to explain more attention effects and to better identify the underlying mechanisms. Spike models of attention (Ardid et al., 2007; Hugues and José, 2010; Buia and Tiesinga, 2008; Wagatsuma et al., 2013) have the problem that they can only account for a few effects (Chap. 3.4.1). We suspect that the reason is an increased complexity in spike models due to the more detailed neuronal processing. This is especially visible in the model of Wagatsuma et al. (2013), which simulates the full cortical microcircuit, but is only able to replicate two attention effects. Even worse, it is impossible to relate these effects to a particular neuronal connection or model mechanism as the network dynamics are far too complex. Therefore, we think the level of mean-rate coded neurons is more suitable for explaining a large amount of attention effects within a single framework.

Secondly, we limit the topic of learning to the learning of object descriptors (Chap. 4.2.4, 5.2.2). We do not cover any other aspects of learning in the visual cortex (Chap. 4.5.4) as this would be a doctoral thesis on its own. We implement for example the learning in a separate stage, and do not address the fact that learning in the cortex is not separated.

Furthermore, the visual cortex learns its representation also on the basis of category information (Sigala et al., 2002). We do not exploit such learning aspects as the precise cortical mechanisms of the required error signal are not well understood (Chap. 4.4.2).

Finally, we constrain ourselves to the modeling of two psychophysical attention phenomena, although many other phenomena exist that may be also worth to model (Chap. 2.5). Even within a phenomenon, multiple variations can exist. For example, OSM encompasses variation with one, two, or four masks, and our model would predict differing attentive processing for them (Chap. 5.5.2). Therefore, it may be very interesting to model more phenomena and to identify the invoked attention mechanisms.

6.2. Conclusion

We elaborated in this doctoral thesis common neuronal mechanisms of visual attention to explain many of its different aspects at once and to provide an implementation prototype for future computer vision systems. Attention has many and seemingly-different aspects, both at neuronal and behavioral level, and thus is classically hard to grasp. We explained here many different aspects of attention with a same set of underlying neuronal mechanisms, allowing a more uniform understanding of attention and a bridging of the gap between its neuronal and behavioral aspects. The neuronal mechanisms were simulated in neuro-computational models, resulting in a neuro-computational framework that explains for the first time various divergent facets of attention at once. During our course of action, we also advanced the state-of-the-art in each field to further deepen our understanding of visual attention in these subfields.

In the first part of the thesis (Chap. 3), we explained the neurophysiological effects of attention, and based on this modeling work, identified the neuronal mechanisms of attention. To explain a broad range of effects at once, we developed a novel microcircuit model by unifying existing models of attention and underpinned it further with neuroanatomical constraints. We identified four neuronal mechanisms as essential for attention: 1) amplification of neurons preferring an attended stimulus, 2) divisive normalization of the response, 3) spatial pooling within the receptive field of a neuron, and 4) suppression from neurons encoding different features or different locations in the surround. With them, the microcircuit model explains a vast range of neurophysiological attention effects: spatial- and feature-based biased competition (Busse et al., 2009; Chelazzi et al., 1998; Lee and

Maunsell, 2010a; MacEvoy et al., 2009; Reynolds et al., 1999), modulation of the contrast response function (Reynolds et al., 2000; Williford and Maunsell, 2006), modulation of the neuronal tuning curve (McAdams and Maunsell, 1999; Martinez-Trujillo and Treue, 2004; Treue and Trujillo, 1999), and modulation of surround suppression (Cavanaugh et al., 2002b; Sundberg et al., 2009). Moreover, it predicts novel attention effects, which we demonstrated by combining the biased competition and surround suppression paradigms.

Afterwards, we model with these neuronal mechanisms behavioral phenomena of attention. For this, we scaled up the microcircuit model to a system-level model of attention (Chap. 4), containing major relevant areas for attentional processing: a lower and a higher visual area (HVA), the frontal eye field (FEF), and the prefrontal cortex (PFC). The model is underpinned with neuroanatomical and neurophysiological properties like reported cell types of higher visual areas and the FEF, receptive field sizes, and synaptic transmission delays. As feature-based and spatial attention operate at system-level partly over different circuits, the underlying amplification and suppression mechanisms exist now both as feature-based and spatial variants in the system-level model. To better understand the role of attention at system-level, we also re-investigated the available neuroscience literature (Chap. 4.2.3). We found that the attentional processing spawns a top-down control network, modulating neuronal activity for the current task. Therefore, we proposed a novel, more general view of attention than the classical view as a selection process: we see attention as a cognitive and holistic control process, tuning the visual system for the task at hand. The control is termed as holistic as it modulates activity in parallel and across the whole visual cortex.

As the first behavioral phenomenon, we investigated the localization of objects (Chap. 4), also known as guided visual search. This phenomenon was chosen as object localization is a very relevant computer vision problem and to demonstrate the robustness of our approach. Attention has been already used in computer vision systems for this task, but often merely as a spatial pre-selection stage (saliency models). This approach has some drawbacks which are avoided by neuro-computational models due to their holistic attention processing, thus we like to propose their usage as a novel alternative. For this, we illustrate again the attentional processing in object localization. However, neuro-computational models have not been demonstrated yet on realistic problems with many objects and large setups, as often found in the computer vision domain. Therefore, we first showed with our novel model that they are also applicable to real problems. For this, we demonstrated our system on a large and realistic setup with 100 objects, 1000 scenes, and three background classes (black, white-noise, real-world). Our model can cope with such a high number of objects as we introduced learned object representations, which then are attentionally modulated. This shows how attention can guide vision with a high number of object categories. Our model achieves the following accuracies to localize objects: It reaches an accuracy of 92% at black backgrounds, at which it was learned (native setup). Errors occur predominantly when the target and distractor are similar, which is reasonable as also humans make errors under these conditions. If the model has to generalize to other backgrounds, the accuracy changes to 71% for white-noise and to 42% for real-world backgrounds respectively. Secondly, we explained the attentional processing in object localization and the roles of the neuronal attention mechanisms in this task. According to our model, attention operates as follows: The target object is encoded in prefrontal cortex, which sends out top-down attention signals to higher visual areas that amplify there all cells encoding features of the target (feature-based attention). Spatial cortices, e.g. the FEF, rely then on this biased activity pattern and select by means of a recurrent processing the target location. We found that the mechanisms have the following roles in this processing: a) The feature-based amplification mechanism performs the enhancement of the target neurons, and thus represents the target in the visual system for the top-down control so other mechanisms can then rely on this biased activity, and b) the feature-based suppression is irrelevant for the behavioral performance in this task, but removes neuronal noise. Neuronal noise denotes here every activity resulting from distractors or background clutter.

As the second phenomenon (Chap. 5), we explained OSM with our attention model to demonstrate that the modeling approach can account for psychophysical data, and to show the link between the seemingly-different phenomena of attention and masking. OSM is explained prominently through three neuronal attention mechanisms: 1) The masking effect, i.e. the impaired recognition of the target stimulus, is explained by feature-based suppression from neurons encoding the mask to neurons encoding the target. 2) In OSM, the neuronal target response has to be maintained after stimulus offset, which is achieved by the amplification mechanism. 3) OSM requires a high number of distractors, denoted as set size effect. This occurs because of a delayed onset of spatial attention, weakening the influence of the underlying spatial amplification mechanism. Our model is one of the first neuro-computational models of OSM that is solidly grounded on neuroanatomical and neurophysiological data. Thus, it allows a much more solid investigation of open research questions than existing models: We found out that reentrant processing might not be essential for OSM, and that the factors mask duration and set size are indeed independent.

Finally, we compared the role of the neuronal mechanisms in each attention aspect to link the different facets of visual attention together (Chap. 6.1). At first, we bridged the gap between neurophysiology and behavior by summarizing the role of the neurophysiologically-founded mechanisms in the two behavioral phenomena (Sec. 6.1.2). We found that the amplification mechanism is strongly relevant for behavior as it encodes the target stimulus of the current task. The suppression is similarly important as it removes noise in the neuronal code and implements a competition between behavioral alternatives. The other mechanisms, spatial pooling and divisive normalization, are less relevant for behavior, at least in the tasks investigated here. Secondly, we analyzed the role of the mechanisms within the two phenomena (Sec. 6.1.1). The two phenomena represent a localization and recognition task respectively. A localization task is characterized through that the object type is given and the object location has to be determined. A recognition task has precisely the opposite task instruction: the object location is given and its type has to be determined. OSM is a recognition task as the location is given by the four-dot mask and the type of the target object has to be determined. Hence, the two task classes have precisely divergent task demands. We found out that our model utilizes diametrically the feature-based and spatial variants of the amplification and suppression mechanisms to meet the divergent tasks demands. This finding shows how attention as cognitive control is able to realize different task classes, and predict how attention in the brain might operate in tasks belonging to these classes. Furthermore, it indicates a broad applicability of our modeling approach as our model is able to simulate both fundamentally different task classes.

To summarize, we explained many different aspects of visual attention with a set of four neuronal mechanisms in a single neuro-computational framework: numerous single-cell recordings, object localization, and OSM. Modeling object localization highlights the robust functionality of the system, while OSM shows its ability to account for psychophysical data. Therefore, our framework's mechanisms allow to explain attention from a neurophysiological, functional, and psychophysical point of view. This explanation power provides strong evidence that these neuronal mechanisms might constitute the general neuronal substrate of visual attention in the brain.

6.3. Outlook

The current work has elaborated the essential neuronal mechanisms of visual attention, and thus, deepened notably our fundamental understanding of attention. Future studies may either continue this computational-neuroscience research by examining the mechanisms of attention in other psychophysical phenomena, or they may develop powerful computer vision systems by using the elaborated mechanisms as a concrete implementation prototype.

In the field of neuroscience, it would be interesting to understand the dynamics of attention in other psychophysical phenomena (Chap. 2). Especially interesting are the ones that have never been simulated by models of attention: crowding, attentional blink to some degree, and multiple object tracking. Such modeling studies would also benefit the computer science community as some of the phenomena constitute typical computer vision problems, e.g. motion perception.

For computer science, the concept of attention as a cognitive, holistic control may be very beneficial for future computer vision systems. For example in object localization, it has at least two theoretical advantages over top-down saliency models as the state-of-the-art approach: parallel localization and recognition, and localization via high-level object representations (Chap. 2.4.3). Therefore, future studies should apply our approach to computer vision problems, to benchmark it on more applications and evaluate its advantages. However, our model is relatively slow compared to computer vision approaches (Chap. 4.5.4), but it can be easily accelerated for an application. The lack of speed results from simulating accurately the neurophysiological responses over time by using ordinary differential equations (ODEs) and the Euler method. This approach needs to evaluate the whole network at every millisecond, requiring for the recognition of a single image 150 - 750 evaluations of the network. In contrast, typical computer vision approaches perform only a single evaluation. We think that the accurate modeling of responses is not necessary in computer vision systems as their primary goal is a correct decision. In our model, a decision relies on the differences in the neuronal responses at a few time points, whereas all other aspects of the responses matter less. Thus, it would be enough to evaluate the network only at a few time points, which would massively speedup the system. A such accelerated version of our system should be fast enough to be applied in the computer vision domain.

According to the holistic attention concept, attention provides a task-dependent control process that tunes the visual system for the current task. Such a control is currently not part of computer vision systems, thus it might be very beneficial to combine it with the best available computer vision systems. Currently, these are deep neuronal networks (LeCun et al., 1998, 2015) as they have shown outstanding performances on the ImageNet challenge, which benchmarks the recognition and localization of objects in real-world scenes (Krizhevsky et al., 2013). Their strength is that they have a very sophisticated representation of visual stimuli at several hierarchy levels, ranging from representing edges at lower levels to objects at higher levels (Zeiler and Fergus, 2014). Attention as holistic control has not been combined with deep networks yet. We expect strong benefits from a combination, given the attention approach's ability to tune the visual system for the current task (Chap. 4.2.3), and given its advantages over the saliency models (Chap. 2.4.3). Aside from these, attention seems to be essential for visual processing as it is involved in almost any visual perception phenomenon (Chap. 2.5). On the other hand, attentive control would also benefit from deep networks as they provide a sophisticated object representation, which is in our model simpler, and thus the main source of errors (Chap. 4.5.4). Therefore, we expect that the combination of both systems may achieve an even better performance than the great performance of deep networks. The best candidates for a combination are convolution networks (LeCun et al., 2015) as their core operations can be mapped into our model: their convolution operation corresponds to our weighted sum within the excitation, and their max-pooling operation corresponds to our spatial pooling mechanism. Furthermore, their representation of visual stimuli is to some extent comparable to the ones in the visual cortex (Khaligh-Razavi and Kriegeskorte, 2014; Kriegeskorte, 2015). After the combination, a system would provide the best of both worlds: a sophisticated representation of the visual stimuli in our world, and an attentive control process tuning the system for the needs of the current task. We think, these two components are the main reasons why the human visual system achieves such an outstanding performance. Therefore, this combination may be one of the best ways to develop future computer vision systems, achieving superior performances.

Appendix

A. Microcircuit model of attention - results with standard parameters

In the following, we will verify that the basic model characteristics and not the free model parameters explain the vast range of data sets (Chap. 3). For this, we fixed all model parameters to the standard values outlined in Tab. 3.1 and re-simulated all experiments. We only treat the input tuning curve as free parameter. This curve determines the feature tuning of neurons in layer 4 and layer 2/3. As the data sets were recorded from different brain areas, the feature tuning can be expected to be variable. We decided to treat it still as free parameter, as the width of the curve differs strongly among data sets, e.g. narrow in V1 (Fig. 3.3f) versus broad in MT (Fig. 3.8a). Likewise the baseline, i.e. the activity to anti-preferred features, deviates as well, e.g. weak in V4 (Fig. 3.7) versus average in MT (Fig. 3.8a).

As the data stems from different brain areas, neurons, and monkeys, a perfect fit cannot be expected with a single set of parameters. For example, feature-based attention amplifies the response in a range from +10% in V1 to +35% in MT (Saenz et al., 2002).

The simulation results show that the model can replicate the main effects in all data sets, with the exception of minor deviations in the surround suppression experiment of Cavanaugh et al. (2002a) and Sundberg et al. (2009) (Fig. A.2c, A.2d). The results are grouped similar as in Chap. 3 into biased competition (Fig. A.1), attentional modulation of the neuronal tuning curve (Fig. A.3), contrast response function (Fig. A.2a,b), and surround suppression (Fig. A.2c - g).

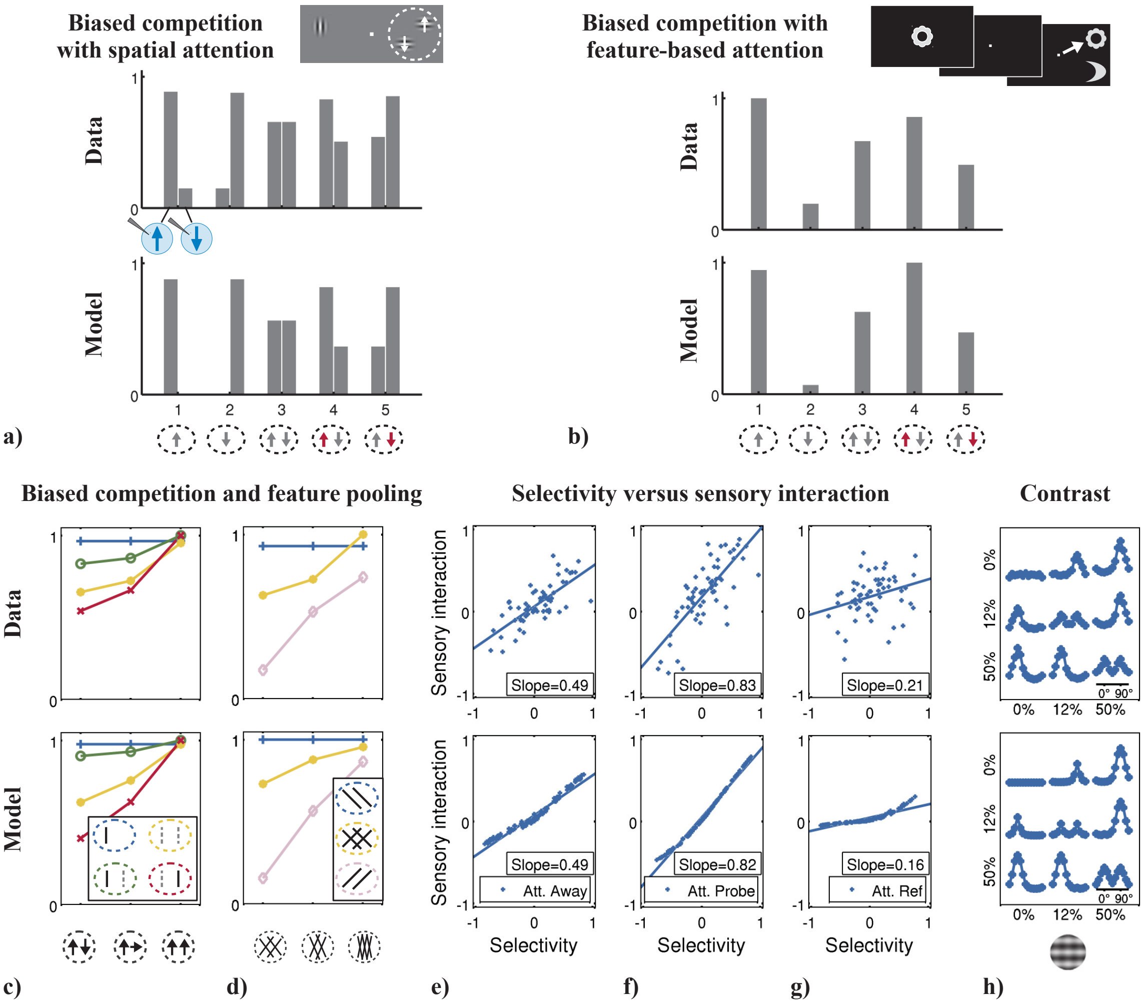


Figure A.1.: Biased competition can be fully replicated with standard parameters. Neurophysiological data (top row) is shown in relation to simulation results (bottom row). The figure refers to Fig. 3.2 - 3.4. **a)** Biased competition with spatial attention: A neuron preferring upward motion shows a lower response to the pair (condition 3) compared to a stimulus alone (condition 1). The response increases if the preferring stimulus is attended (condition 4) or decreases if the anti-preferred is attended (condition 5). **b)** Biased competition with feature-based attention results in the same effects. **c)** The competitive interactions depend on the similarity of the two spatially disjunctive stimuli. Biased competition occurs with two different stimuli (left) and feature pooling with two equal stimuli (right). **d)** The same similarity dependency as in c) can be observed for overlapping gratings in data from V1 and the model. **e-g)** Effect of attention on the relationship between selectivity (probe - reference) and sensory-interaction (pair - reference). Significant effects are: the slope does not differ from 0.5 in the attend away condition (e), is greater than 0.5 in attend probe condition (f), and is less than 0.5 in a attend reference condition (g). **h)** Contrast dependency of cross-orientation suppression. The response to a pair of stimuli is lower as to a single stimulus, illustrating cross-orientation suppression. In data from V1 and in the model, the suppression and the population response depends on the contrast differences between the two stimuli. For similar contrasts, the population response shows two peaks, whereas it shows only a single peak for large differences (i.e. 50% and 12%), as the response to the lower contrasted stimulus is completely suppressed.

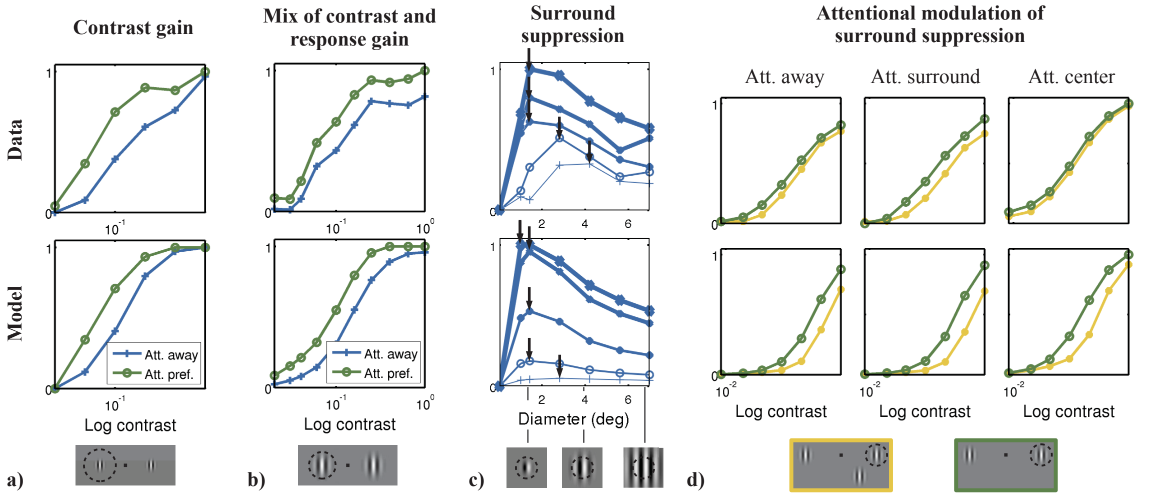


Figure A.2.: Attentional modulation of the contrast response function (a, b) and of surround suppression (c, d). Effects in a), b) can be nearly replicated with standard parameters, effects in c) and d) partly. Neurophysiological data (top row) is shown in relation to simulation results (bottom row). The figure refers to Fig. 3.5, 3.6, and 3.9. **a)** Contrast gain denotes the effect that the response at lower contrasts is significantly stronger increased by attention, resulting in a left-wards shift of the contrast response function (cRF). **b)** A mix of response and contrast gain results in an leftwards and upwards shift of the cRF. The response gain at high contrasts observable in the data is only partially replicated in the model due to a too weak suppression from neurons in the close surround. This suppression is calibrated based on the data of Cavanaugh et al. (2002a) (A.2c), which is optimally fitted by a scaling factor $v^{\text{SUR}} = 0.5$, whereby the free parameter model uses $v^{\text{SUR}} = 1.0$ to fit the response gain data. In the Cavanaugh et al. (2002a) data, such a stronger suppression from the close surround would result in a too swift decrease of the response after the stimulus exceeds the receptive field, i.e. after the peak of the response curve. **c)** Surround suppression without attention: The response to increased gratings shows significant surround suppression after the stimulus exceeds the receptive field border (black arrows). The peak amplitude differs dependent on the stimulus contrast (thickest line denotes highest contrast), which is roughly replicated by the standard parameter model. Furthermore, the receptive field size increases significantly with lower contrasts. Our results show this effect only marginally. The main reason for this deviation is the change of p^{SUR} from 2.0 to 1.0. This parameter controls non-linearly the influence of the presynaptic firing rate on the surround suppression. A value of 2.0 decreases the surround suppression for lower rates induced by lower contrasts, which explains our data fits (Sec. 3.3.4). Contrary, the non-linearity was disabled by setting p^{SUR} to 1.0 in the standard parameter set as it fits optimally four other data sets (Fig. 3.3b, 3.6, 3.7, 3.8a). The difference between these data sets and the current one is the large regular stimulus in the current one. This stimulus would drive optimally contour linkage effects (Gilbert, 1998), which would enhance the response and so decrease suppression. Therefore, we speculate that the increase of receptive fields is due to contour linkage effects which is beyond the grasp of the attention mechanisms in our model.

Figure A.2 continued.: d) Attentional modulation of surround suppression: In attend away, adding the surround stimulus suppresses the response over all contrasts, whereby the suppression is proportionally larger at lower contrasts. Attention to the surround increases this suppression, while attention to the center decreases it. All these significant effects can be replicated. Yet, our result deviates notably from the data because the surround suppression is too strong (yellow curves). The model with free parameters fits the data via a weak suppression $v^{\text{SUR}} = 0.55, p^{\text{SUR}} = 2.5$, on the contrary the standard model uses $v^{\text{SUR}} = 0.5, p^{\text{SUR}} = 1.0$. The non-linearity ($p^{\text{SUR}} = 2.5$) has the side effect to weaken the suppression, so it is much weaker than in the standard model with disabled non-linearity ($p^{\text{SUR}} = 1.0$). The non-linearity is disabled in it as it allows to fit more data sets (Fig. 3.3b, 3.6, 3.7, 3.8a).

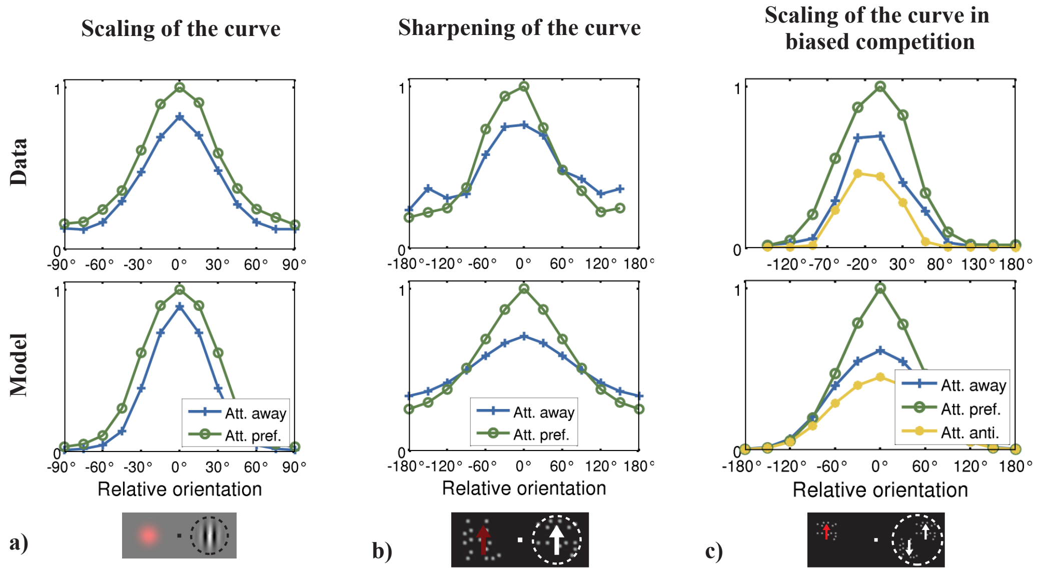


Figure A.3.: Attention modulation of neuronal tuning curves can be fully replicated with standard parameters. Neurophysiological data (top row) is shown in relation to simulation results (bottom row). The figure is illustrated identically as Fig. 3.7 - 3.8. **a)** Scaling of a tuning curve by spatial attention. The authors reported two significant results, which are also present in our results: The whole tuning curve is scaled, and is slightly, but significantly shifted upwards by attention. The width and the preferred tuning direction are not affected by attention. **b)** Sharpening of the tuning curve is caused by feature-based attention, resulting in a significantly increased response at the preferred feature (0°) and in a significantly decreased response at anti-preferred ones (±180°). c) On the contrary in biased competition, feature-based attention leads to a scaling of the tuning curve and sharpening cannot be observed.

B. Object localization with a neuro-computational model of visual attention

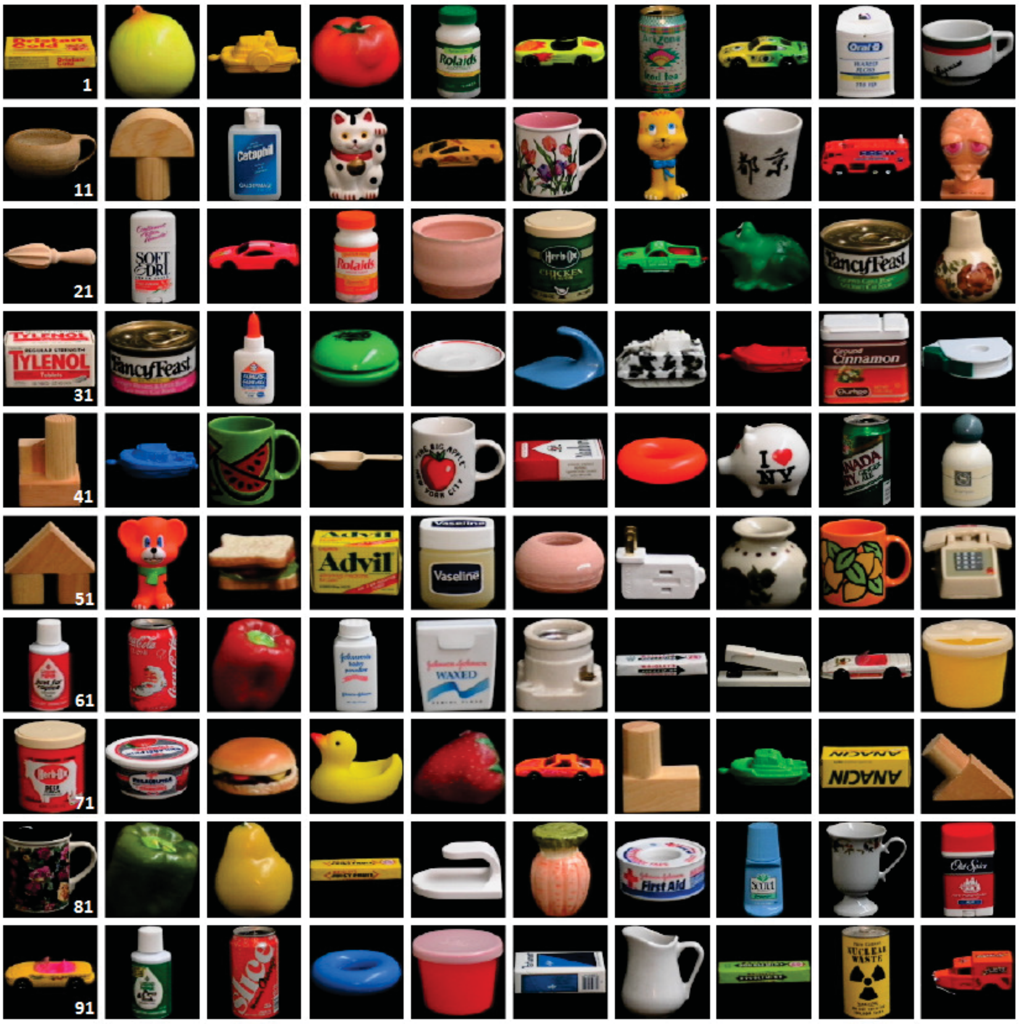


Figure B.1.: All 100 objects of the COIL-100 database (Nene et al., 1996). The objects are ordered according to their ID, displayed in the first column.

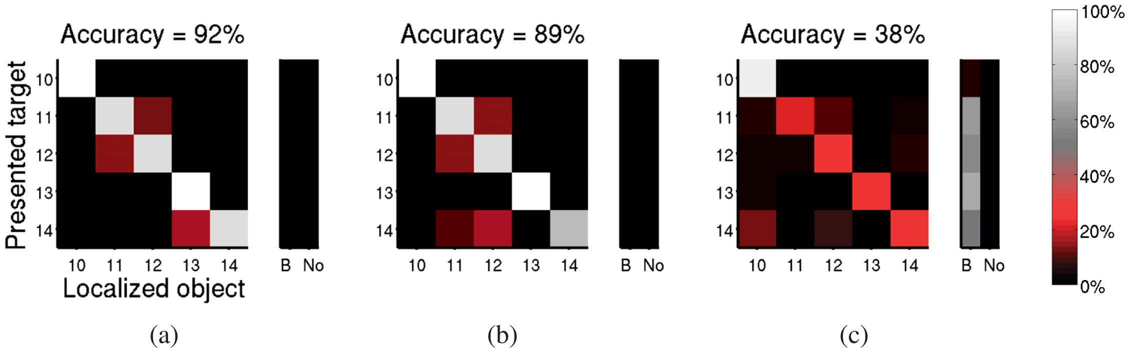


Figure B.2.: Confusion matrices (Sokolova and Lapalme, 2009) for the 5 object test sets with black **(a)**, white-noise **(b)**, and real-world backgrounds **(c)**. Low values under 50% are illustrated in red to show mislocalizations. For all matrices, the x-axis denotes the localized object or two special cases. The case "B" indicates that the background was selected by a saccade. The case "No" indicates that no particular location was selected because the system did not execute a saccade. The axis numbers refer to the objects IDs in the COIL-100 database (Fig. B.1).

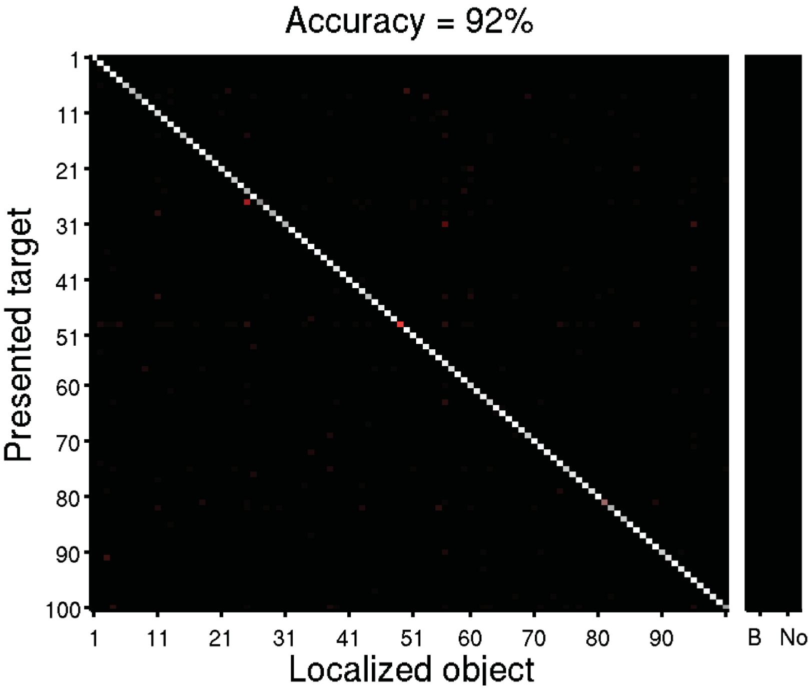


Figure B.3.: Confusion matrix for the 100 object test sets with black backgrounds. The matrix is illustrated as in Fig. B.2.

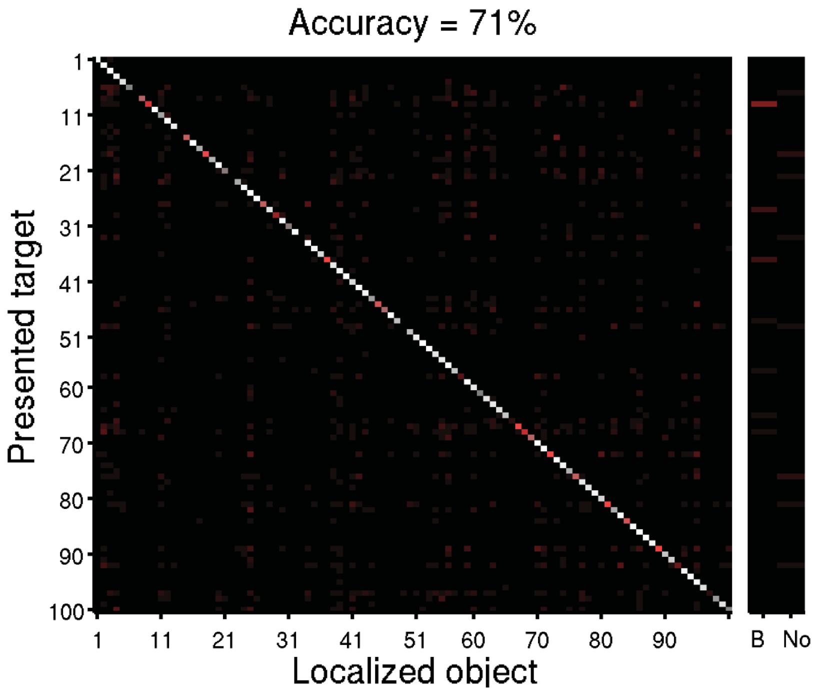


Figure B.4.: Confusion matrix for the 100 object test sets with white-noise backgrounds. The matrix is illustrated as in Fig. B.2.

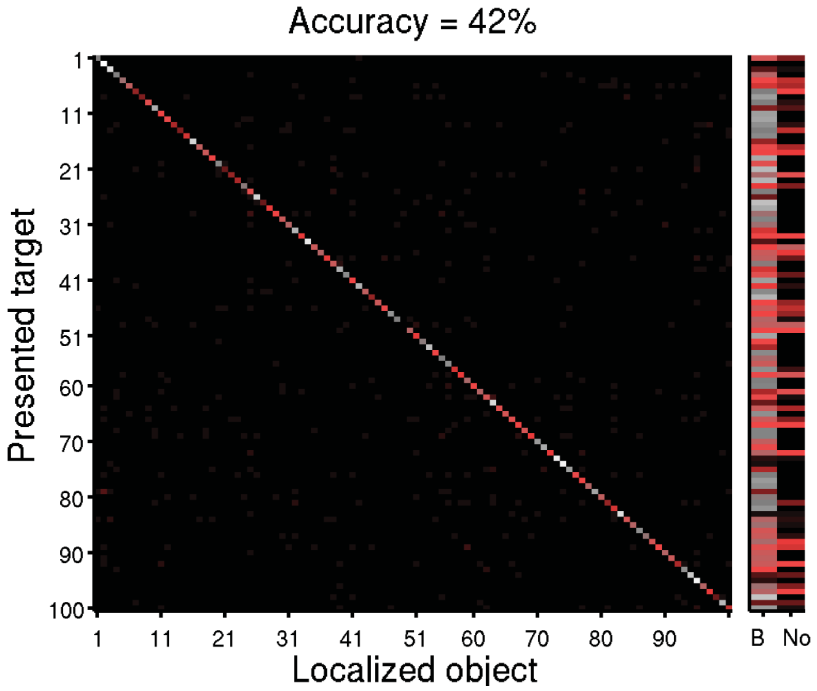


Figure B.5.: Confusion matrix for the 100 object test sets with real-world backgrounds. The matrix is illustrated as in Fig. B.2.

C. Alpha function to simulate spatiotemporal receptive fields

This appendix section describes the concept of an alpha-function, developed within the research group of Prof. Hamker by Tobias Höppner, Abbas Al Ali, and Fred Hamker. The alpha function describes a time variant function, which is used to model the temporal processing in the early stages of the visual system, i.e. in the retina and lateral geniculate nucleus (LGN). The following section was provided by Tobias Höppner.

C.1. Time variant function in the retina and the LGN

Processing of light in the retina involves a number of layers, containing photoreceptors and neuron-like cells (Wässle, 2004). The mammalian retina contains upward of 50 distinct functional elements, each carrying out a specific task (Masland, 2001). The retina is privileged because of its physical location, the distinct morphology of its neurons and its accessibility of input and output.

Here, we work out a simple model based on synaptic transmissions, which describes part of retina processing (Borst and Euler, 2011). Photoreceptors emit electrical discharges to transduce the energy of the incoming light into electrical signals. These discharges have the typical form of decaying particles, which can be easily modeled by a first-order ordinary differential equation (ODE). After the emission, the discharges are converted by neuron-like cells to spikes and these are transmitted via synapses. This synaptic transmission can be described by a similar differential equation than for the decaying particles. These equations have a particular decaying time course and are referred as alpha functions later.

After the retina, the next structures in the processing are the LGN and the primary visual cortex (V1). It is widely assumed that the bipolar cells in the retina form a Gaussian-shaped receptive field which is transmitted to the ganglion cells in the LGN, because this difference of Gaussian spatial structure is found in the LGN (Wiesel and Hubel, 1966) as well as V1 (Hubel and Wiesel, 1962). However, receptive fields in these areas have also a varying time component, which may result from the processing in the retina. Such kind of receptive fields are called spatiotemporal receptive fields (Fig. C.1), and have been found in the LGN (Cai et al., 1997) and V1 (DeAngelis et al., 1993).

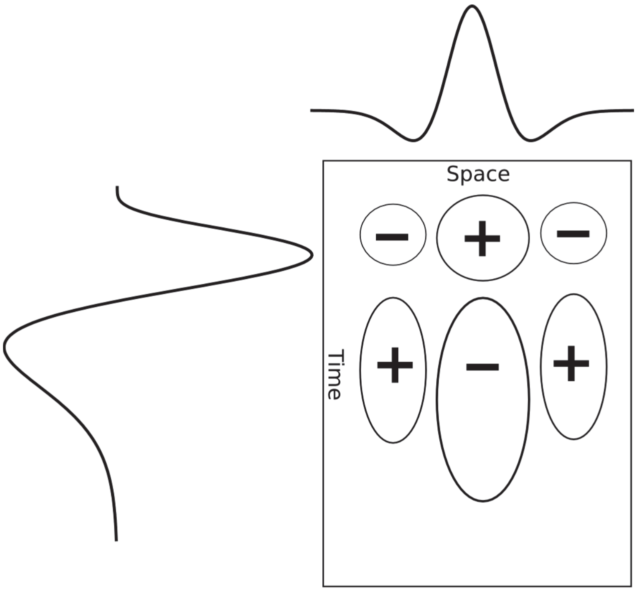


Figure C.1: A spatiotemporal receptive field, which has been found in LGN (Cai et al., 1997) or V1 (DeAngelis et al., 1993). The figure is a schematic illustration, adopted from Adelson and Bergen (1985).

C.2. Modeling spatiotemporal receptive fields

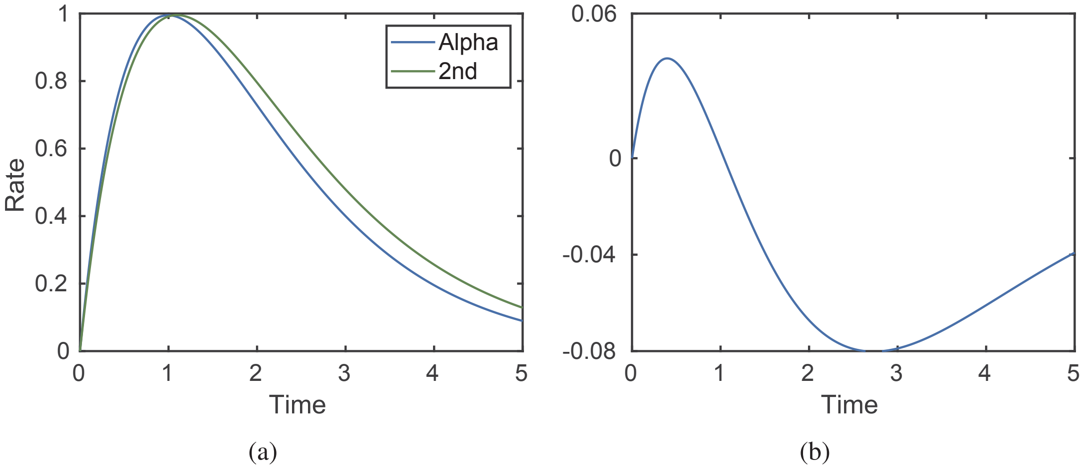
A spatiotemporal receptive field in LGN consists of a spatial component and a temporal component (Fig. C.1): The spatial component is typically modeled by a difference of Gaussian approach. The temporal component is mostly approximated by a biphasic function, which describes the development in time of the field. From a mathematically point of view, it is possible to construct a biphasic function out of two alpha-functions. An alpha function has a distinct temporal shape, which is shown in Fig. C.3b.

Such an alpha function can be dynamically established by a coupled ODE system as shown by the equations C.1 and C.2. One such equation block outputs one alpha function. By varying the time course (parameter τ_s) or the maximum of the function (parameter $\alpha_{\max}$), it is possible to adapt the function to a wide range of physiological data.

$$\tau_s \frac{d\,\alpha}{d\,t} = e\,\alpha_{\max}\,z - \alpha \tag{C.1}$$

$$\tau_s \frac{d\,z}{d\,t} = -z \tag{C.2}$$

This model tries to extract the time course of the field components out of the decaying strength of the input function. As a biphasic function is a function that is the difference of two alpha functions, it is possible to construct the time course with two evaluation blocks.



(a)

(b)

Figure C.2.: Two alpha functions **(a)** and a biphasic function **(b)**. The biphasic function is the difference between the two alpha functions.

For one biphasic function, one needs to evaluate two alpha function blocks, each with a slightly different τ-value. The result of such a process is depicted in Fig. C.2. The resulting biphasic function can now be multiplied with the spatial component of the receptive field. After the multiplication, the field exhibits the intended temporal behavior.

Derivation of an alpha function

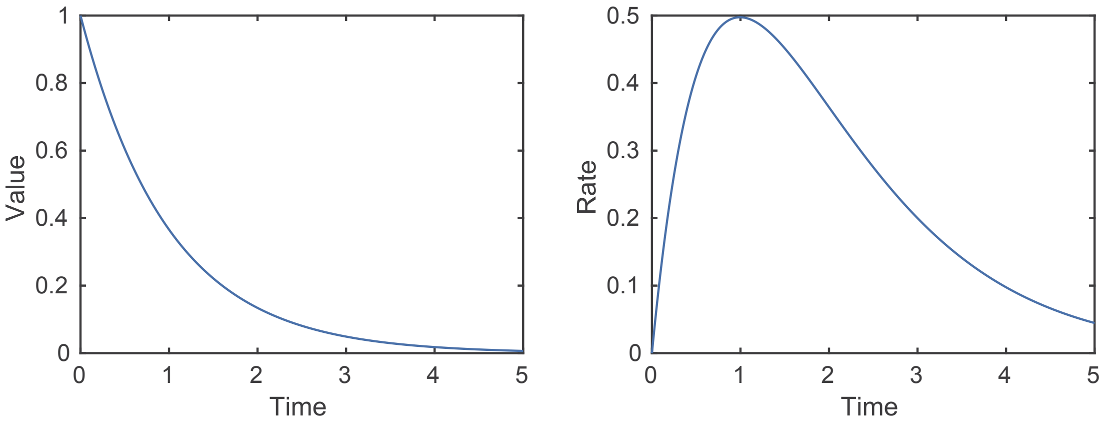
From a synaptic transmission point of view, an alpha-function (Fig. C.3b) describes the probability of a postsynaptic channel to be open (P_s):

$$P_s = \frac{P_{\max}\, t\, e^{1-\frac{t}{\tau_s}}}{\tau_s} \tag{C.3}$$

Now, we will derive a system of ordinary differential equations (ODEs) from an alpha function. First, one has to differentiate the equation with respect to t:

$$\frac{d\,P_s}{d\,t} = \frac{d}{d\,t}\,\frac{P_{\max}\, t\, e^{1-\frac{t}{\tau_s}}}{\tau_s} \tag{C.4}$$

$$\frac{d\,P_s}{d\,t} = \frac{P_{\max}\, e^{1-\frac{t}{\tau_s}}}{\tau_s} - \frac{P_{\max}\, t\, e^{1-\frac{t}{\tau_s}}}{\tau_s^2} \tag{C.5}$$



(a) (b)

Figure C.3.: Helper function z **(a)** and the resulting alpha-function **(b)**. The graphs are obtained with τ_s = 100 ms and $P_{\max}$ = 0.5.

By substituting the functional relation of P_s into this equation, one obtains:

$$\frac{d\,P_s}{d\,t} = \frac{P_{\max}\,e^{1-\frac{t}{\tau_s}}}{\tau_s} - \frac{P_s}{\tau_s} \tag{C.6}$$

We multiply the equation with τ_s:

$$\tau_s\left(\frac{d\,P_s}{d\,t}\right) = \tau_s\left(\frac{P_{\max}\,e^{1-\frac{t}{\tau_s}}}{\tau_s} - \frac{P_s}{\tau_s}\right) \tag{C.7}$$

The equation is simplified to:

$$\tau_s\,\frac{d\,P_s}{d\,t} = P_{\max}\,e^{1-\frac{t}{\tau_s}} - P_s \tag{C.8}$$

With the definition of a helper function z (Fig. C.3a):

$$z = e^{-\frac{t}{\tau_s}} \tag{C.9}$$

The function z describes a discharge function after an arrival of a presynaptic spike. Thus, z will be 1 when a spike arrive, modeled by z(t=0)=1.

We get the primary ODE for $\mathrm{P_s}$:

$$\tau_s \frac{d\,P_s}{d\,t} = e\,P_{\max}\,z - P_s \tag{C.10}$$

Therefore, the basic building block of a generic alpha-function is a coupled ODE in the following form:

$$\tau_s \frac{d\,\alpha}{d\,t} = e\,\alpha_{\max}\,z - \alpha \tag{C.11}$$

$$\tau_s \frac{d\,z}{d\,t} = -z \tag{C.12}$$

Bibliography

Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A*, 2(2):284–299.

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

Allman, J., Miezin, F., and McGuinness, E. (1985). Direction- and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception*, 14(2):105–26.

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupe, J.-M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *J Neurosci*, 22(19):8633–46.

Anton-Erxleben, K. and Carrasco, M. (2013). Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat Rev Neurosci*, 14(3):188–200.

Anton-Erxleben, K., Stephan, V. M., and Treue, S. (2009). Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cerebral cortex*, 19(10):2466–78.

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.

Ardid, S., Wang, X.-J., and Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *J Neurosci*, 27(32):8486–95.

Argyropoulos, I., Gellatly, A., Pilling, M., and Carter, W. (2013). Set size and mask duration do not interact in object-substitution masking. *J Exp Psychol Hum Percept Perform*, 39(3):646–61.

Aryananda, L. (2002). Recognizing and remembering individuals: online and unsupervised face recognition for humanoid robot. In *Proc IEEE/RSJ Conf Intelligent Robots and Systems 2002 - IROS2002*, pages 1202–1207.

Ashby, F. G. and O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends Cogn Sci*, 9(2):83–9.

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 - Second Edition*. John Wiley & Sons.

Bair, W., Cavanaugh, J. R., and Movshon, J. A. (2003). Time course and time-distance relationships for surround suppression in macaque V1 neurons. *J Neurosci*, 23(20):7690–701.

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.

Bayerl, P. and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Comput*, 16(10):2041–66.

Bayerl, P. and Neumann, H. (2007a). A neural model of feature attention in motion perception. *Bio Sys*, 89(1-3):208–15.

Bayerl, P. and Neumann, H. (2007b). Disambiguating visual motion by form-motion interaction-a computational model. *Int J Comput Vis*, 72(1):27–45.

Beck, C. and Neumann, H. (2010). Interactions of motion and form in visual cortex - A neural model. *J Physiol (Paris)*, 104(1-2):61–70.

Beck, D. M. and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res*, 49(10):1154–65.

Beuth, F. and Hamker, F. H. (2015a). A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Res*, 116(Part B):241–57.

Beuth, F. and Hamker, F. H. (2015b). Attention as cognitive, holistic control of the visual system. In Hammer, B., Martinetz, T., Schleif, F.-M., and Villmann, T., editors, *Proc Workshop New Challenges in Neural Computation 2015 - NCNC 2015, Machine Learning Reports 03/2015*, pages 133–140.

Beuth, F., Jamalian, A., and Hamker, F. H. (2014). How Visual Attention and Suppression Facilitate Object Recognition? In Wermter, S., Weber, C., Duch, W., Honkela, T., Koprinkova-Hristova, P., Magg, S., Palm, G., and Villa, A., editors, *Proc 24th Int Conf Artificial Neural Networks - ICANN 2014, Lecture Notes in Computer Science 8681*, pages 459–466.

Beuth, F., Wiltschut, J., and Hamker, F. H. (2010). Attentive Stereoscopic Object Recognition. In Hammer, B., Schleif, F.-M., and Villmann, T., 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.

Bisley, J. W. and Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annu Rev Neurosci*, 33:1–21.

Björkman, M. and Eklundh, J. (2006). Vision in the real world: Finding, attending and recognizing objects. *Int J Imag Syst Tech*, 16(5):189–208.

Bonin, V., Mante, V., and Carandini, M. (2005). The suppressive field of neurons in lateral geniculate nucleus. *J Neurosci*, 25(47):10844–56.

Borji, A., Ahmadabadi, M. N., and Araabi, B. N. (2009). Cost-sensitive learning of top-down modulation for attentional control. *Mach Vis Appli*, 22(1):61–76.

Borji, A., Ahmadabadi, M. N., Araabi, B. N., and Hamidi, M. (2010). Online learning of task-driven object-based visual attention control. *Image Vision Comput*, 28:1130–1145.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans Pattern Anal Mach Intell*, 35(1):185–207.

Born, R. T. and Bradley, D. C. (2005). Structure and function of visual area MT. *Annu Rev Neurosci*, 28:157–89.

Borst, A. and Euler, T. (2011). Seeing Things in Motion: Models, Circuits, and Mechanisms. *Neuron*, 71(6):974–994.

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

Boyer, J. and Ro, T. (2007). Attention attenuates metacontrast masking. *Cognition*, 104(1):135–49.

Boynton, G. M. (2009). A framework for describing the effects of attention on visual responses. *Vision Res*, 49(10):1129–43.

Breitmeyer, B. G. (1984). *Visual Masking: An Integrative Approach*. Oxford University Press.

Breitmeyer, B. G. and Ogmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. Oxford University Press.

Bridgeman, B. (1978). Distributed sensory coding applied to simulations of iconic storage and metacontrast. *B Math Biol*, 40(1973):605–623.

Bridgeman, B. (2007). Common-onset masking simulated with a distributed-code model. *Adv Cogn Psychol*, 3(1-2):33–40.

Broadbent, D. E. (1958). *Perception and communication*. Pergamon Press.

Brosch, T. and Neumann, H. (2014). Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Net*, 54:11–6.

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

Bruce, N. D. B. and Tsotsos, J. K. (2011). Visual representation determines search difficulty: explaining visual search asymmetries. *Front Comp Neurosci*, 5:33.

Bruce, N. D. B. and Tsotsos, J. K. (2012). Attention in Stereo Vision : Implications for Computational Models of Attention. In *Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts*, pages 65–88.

Buia, C. I. and Tiesinga, P. H. (2008). Role of interneuron diversity in the cortical microcircuit for attention. *J Neurophysiol*, 99(5):2158–82.

Bullier, J. (2001). Integrated model of visual processing. *Brain Res Rev*, 36(2-3):96–107.

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.

Busse, L. (2006). *Effects of Selective Attention on Sensory Processing of Visual Motion*. Doctoral dissertation, Georg-August-Universität Göttingen, Germany.

Busse, L., Wade, A. R., and Carandini, M. (2009). Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, 64(6):931–42.

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.

Cai, D., DeAngelis, G. C., and Freeman, R. D. (1997). Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *J Neurophysiol*, 78(2):1045–61.

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

Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat Rev Neurosci*, 13(1):51–62.

Carpenter, G. a., Martens, S., and Ogas, O. J. (2005). Self-organizing information fusion and hierarchical knowledge discovery: A new framework using ARTMAP neural networks. *Neural Netw*, 18:287–295.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Res*, 51(13):1484–525.

Cass, J. R. and Spehar, B. (2005). Dynamics of collinear contrast facilitation are consistent with long-range horizontal striate transmission. *Vision Res*, 45(21):2728–39.

Cavanagh, P. (1992). Attention-based motion perception. *Science*, 257(5076):1563–1565.

Cavanagh, P. and Alvarez, G. a. (2005). Tracking multiple targets with multifocal attention. *Trends Cogn Sci*, 9(7):349–354.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J Neurophysiol*, 88(5):2530–46.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *J Neurophysiol*, 88(5):2547–56.

Chakravarthi, R. and Cavanagh, P. (2009). Recovery of a crowded object by masking the flankers: determining the locus of feature integration. *J Vis*, 9(10):4.

Chapman, B. M., Gropp, W. D., Kumaran, K., and Müller, M. S. (2011). *OpenMP in the Petascale Era: 7th International Workshop on OpenMP, IWOMP 2011, Chicago, Il, USA, June 13-15, 2011, Proceedings*, volume 6665. Springer Science & Business Media.

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

Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiol*, 80(6):2918–40.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral cortex*, 11(8):761–72.

Chen, Y. and Martinez-Conde, S. (2008). Task difficulty modulates the activity of specific neuronal populations in primary visual cortex. *Nat Neurosci*, 11(8):974–982.

Chikkerur, S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a Bayesian inference theory of attention. *Vision Res*, 50(22):2233–47.

Chinellato, E. (2008). *Visual Neuroscience of Robotic Grasping*. Doctoral dissertation, Universitat Jaume I, Castellón de la Plana, Valencia, Spain.

Cohen, J. Y., Heitz, R. P., Woodman, G. F., and Schall, J. D. (2009). Neural basis of the set-size effect in frontal eye field: timing of attention during visual search. *J Neurophysiol*, 101(4):1699–1704.

Compte, A. and Wang, X. (2006). Tuning curve shift by attention modulation in cortical neurons: a computational study of its mechanisms. *Cerebral Cortex*, 16(6):761–78.

Congdon, P. (2007). *Bayesian Statistical Modelling - Second Edition*. John Wiley & Sons.

Corchs, S. and Deco, G. (2002). Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. *Cerebral Cortex*, 12:339–348.

Crouzet, S. M., Overgaard, M., and Busch, N. a. (2014). The fastest saccadic responses escape visual masking. *PLoS ONE*, 9(2):e87418.

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

Danilova, M. and Bondarko, V. (2007). Foveal contour interactions and crowding effects at the resolution limit of the visual system. *J Vis*, 7(2):25.

Darrell, T., Gordon, G., Harville, M., and Woodfill, J. (2000). Integrated person tracking using stereo, color, and pattern detection. *Int J Comput Vis*, 37(2):175–185.

David, S. V., Hayden, B. Y., Mazer, J. a., and Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, 59(3):509–21.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *J Neurophysiol*, 69(4):1118–35.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends Neurosci*, 18(10):451–8.

Deco, G., Pollatos, O., and Zihl, J. (2002). The time course of selective visual attention: theory and experiments. *Vision Res*, 42(27):2925–2945.

Deco, G. and Rolls, E. T. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res*, 44(6):621–642.

Deco, G. and Rolls, E. T. (2005). Attention, short-term memory, and action selection: A unifying theory. *Prog Neurobiol*, 76:236–256.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Phil Trans R Soc B*, 353(1373):1245–55.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective attention. *Annu Rev Neurosci*, 18:193–222.

Deubel, H., Schneider, W. X., and Bridgeman, B. (1996). Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vision Res*, 36(7):985–996.

Di Lollo, V., Enns, J. T., and Rensick, R. a. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *J Exp Psychol Gen*, 129(4):481–507.

Di Lollo, V., Enns, J. T., and Rensink, R. a. (2002). Object substitution without reentry? *J Exp Psychol Gen*, 131(4):594–596.

Douglas, R. J. and Martin, K. a. C. (2004). Neuronal circuits of the neocortex. *Annu Rev Neurosci*, 27:419–51.

Dow, B., Snyder, A., Vautin, R., and Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Exp Brain Res*, 44:213–228.

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.

Duffy, C. J. and Wurtz, R. H. (1997). Medial superior temporal area neurons respond to speed patterns in optic flow. *J Neurosci*, 17(8):2839–2851.

Duncan, J. (1984). Selective attention and the organization of visual information. *J Exp Psychol Gen*, 113(4):501–517.

Dux, P. E., Visser, T. A. W., Goodhew, S. C., and Lipp, O. V. (2010). Delayed Reentrant Processing Impairs Visual Awareness: An Object-Substitution-Masking Study. *Psychol Sci*, 21(9):1242–1247.

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychol Rev*, 70(3):193–242.

Egly, R., Driver, J., and Rafal, R. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *J Exp Psychol Gen*, 123(2):161–177.

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalogr Clin Neurophysiol*, 99(3):225–34.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., and König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biol Cybern*, 93(1):79–90.

Elazary, L. and Itti, L. (2010). A Bayesian model for efficient visual search and recognition. *Vision Res*, 50(14):1338–1352.

Emrich, S. M., Burianová, H., and Ferber, S. (2011). Transient perceptual neglect: visual working memory load affects conscious object processing. *J Cogn Neurosci*, 23(10):2968–82.

Enns, J. T. (2004). Object substitution and its relation to other forms of visual masking. *Vision Res*, 44(12):1321–31.

Enns, J. T. and Di Lollo, V. (1997). Object Substitution: A New Form of Masking in Unattended Visual Locations. *Psychol Sci*, 8(2):135–139.

Falkner, A. L., Krishna, B. S., and Goldberg, M. E. (2010). Surround suppression sharpens the priority map in the lateral intraparietal area. *J Neurosci*, 30(38):12787–97.

Fallah, M., Stoner, G. R., and Reynolds, J. H. (2007). Stimulus-specific competitive selection in macaque extrastriate visual area V4. *Proc Natl Acad Sci USA*, 104(10):4165–9.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47.

Ferri, C., Hernández-Orallo, J., and Salido, M. (2003). Volume Under the ROC Surface for Multi-class Problems. Exact Computation and Evaluation of Approximations. Technical report.

Filipe, S. and Alexandre, L. a. (2013). From the human visual system to the computational models of visual attention: a survey. *Artif Intelli Rev*, pages 1–47.

Fischer, B. and Boch, R. (1985). Peripheral attention versus central fixation: modulation of the visual activity of prelunate cortical cells of the rhesus monkey. *Brain Res*, 345(1):111–23.

Fix, J., Schroll, H., Anton-Erxleben, K., Womelsdorf, T., Treue, S., and Hamker, F. H. (2010). Influence of spatial attention on the receptive field shape of neurons in monkey area MT. In *Cinquième conférence plénière Française de Neurosciences Computationnelles, Neurocomp 2010*, pages 147–152.

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

Földiák, P. (1991). Learning Invariance from Transformation Sequences. *Neural Comput*, 3(2):194–200.

Francis, G. (2000). Quantitative theories of metacontrast masking. *Psychol Rev*, 107(4):768–787.

Francis, G. and Hermens, F. (2002). Comment on "Competition for consciousness among visual events: The psychophysics of reentrant visual processes" (Di Lollo, Enns & Rensink, 2000). *J Exp Psychol Gen*, 131(4):590–593.

Franconeri, S. L., Alvarez, G. a., and Cavanagh, P. (2013). Flexible cognitive resources: competitive content maps for attention and memory. *Trends Cogn Sci*, 17(3):134–41.

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

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In Saitta, L., editor, *Proc 13th Int Conf Machine Learning - ICML 1996*, pages 148–156.

Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Doctoral dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn and Fraunhofer-Institut für Autonome Intelligente Systeme, Germany.

Frintrop, S., Nüchter, A., Surmann, H., and Hertzberg, J. (2004). Saliency-based object recognition in 3D data. In *Proc IEEE/RSJ Conf Intelligent Robots and Systems 2004 - IROS2004*, pages 2167–2172.

Frintrop, S., Rome, E., and Christensen, H. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM TAP*, 7(1):1–39.

Fritz, G., Seifert, C., and Paletta, L. (2005). Attentive object detection using an information theoretic saliency measure. In Paletta, L., Tsotsos, J., Rome, E., and Humphreys, G., editors, *Proc 2nd Int Workshop Attention and Performance in Computational Vision - WAPCV 2004*, pages 29–41.

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Curr Opin Neurobiol*, 17(4):437–455.

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

Gao, D. and Vasconcelos, N. (2005a). Discriminant Saliency for Visual Recognition. In *Adv Neural Inf Process Syst 17 - NIPS 2004*, pages 481–488.

Gao, D. and Vasconcelos, N. (2005b). Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proc IEEE Conf Computer Vision and Pattern Recognition 2005 - CVPR 2005*, pages 282–287.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2009). *Cognitive Neuroscience: the Biology of the Mind*. W.W. Norton.

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

Gerstner, W. and Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press.

Gescheider, G. A. (1997). *Psychophysics: the Fundamentals*. Psychology Press.

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

Gilbert, C. D. (1998). Adult cortical dynamics. *Physiol Rev*, 78(2):467–85.

Goodhew, S. C., Pratt, J., Dux, P. E., and Ferber, S. (2013). Substituting objects from consciousness: A review of object substitution masking. *Psychon Bull Rev*, 20(5):859–77.

Grefkes, C. and Fink, G. R. (2005). The functional organization of the intraparietal sulcus in humans and monkeys. *J Anat*, 207(1):3–17.

Gregoriou, G., Gotts, S., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science*, 324(5931):1207–10.

Grossberg, S. (2013). Adaptive Resonance Theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw*, 37:1–47.

Grossberg, S. and Versace, M. (2008). Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Res*, 1218:278–312.

Haider, B., Duque, A., Hasenstaub, A. R., and McCormick, D. a. (2006). Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *J Neurosci*, 26(17):4535–45.

Hamker, F. H. (1999). *Visuelle Aufmerksamkeit und lebenslanges Lernen im Wahrnehmungs-Handlungs-Zyklus*. Doctoral dissertation, Technische Universität Ilmenau, Germany.

Hamker, F. H. (2004a). A dynamic model of how feature cues guide spatial attention. *Vision Res*, 44(5):501–521.

Hamker, F. H. (2004b). Predictions of a model of spatial attention using sum-and max-pooling functions. *Neurocomputing*, 56:329–343.

Hamker, F. H. (2005a). 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. (2005b). 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. (2005c). Modeling attention: From computational neuroscience to computer vision. In Paletta, L., Tsotsos, J., Rome, E., and Humphreys, G., editors, *Proc 2nd Int Workshop Attention and Performance in Computational Vision - WAPCV 2004*, pages 118–132.

Hamker, F. H. (2006). Modeling feature-based attention as an active top-down inference process. *Bio Sys*, 86(1-3):91–9.

Hamker, F. H. (2007). The mechanisms of feature inheritance as predicted by a systems-level model of visual attention and decision making. *Adv Cogn Psych*, 3:111–123.

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.

Hamker, F. H., Zirnsak, M., Calow, D., and Lappe, M. (2008). The peri-saccadic perception of objects and space. *PLoS Comput Biol*, 4(2):e31.

Hamker, F. H., Zirnsak, M., Ziesche, A., and Lappe, M. (2011). Computational models of spatial updating in peri-saccadic perception. *Phil Trans R Soc B*, 366(1564):554–71.

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.

He, S., Cavanagh, P., and Intriligator, J. (1997). Attentional resolution. *Trends Cogn Sci*, 1(3):115–121.

Hearst, M., Dumais, S., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intell Syst Appl*, 13(4):18–28.

Hegdé, J. and Van Essen, D. (2003). Strategies of shape representation in macaque visual area V2. *Vis Neurosci*, 20(3):313–328.

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*. Doctoral dissertation, Swiss Federal Institute of Technology (ETH Zurich), Switzerland.

Herzog, M. H. and Koch, C. (2001). Seeing properties of an invisible object: feature inheritance and shine-through. *Proc Natl Acad Sci USA*, 98(7):4271–4275.

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.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol*, 195(1):215.

Hubel, D. and Wiesel, T. (1974). Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *J Comp Neurol*, 158:295–306.

Hubel, D., Wiesel, T., and Stryker, M. (1978). Anatomical demonstration of orientation columns in macaque monkey. *J Comp Neurol*, 177:361–379.

Hubel, D. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, 160:106–154.

Hugues, E. and José, J. V. (2010). A unified and quantitative network model for spatial attention in area V4. *J Physiol*, 104(1-2):84–90.

Hunt, J. J., Bosking, W. H., and Goodhill, G. J. (2011). Statistical structure of lateral connections in the primary visual cortex. *Neural Syst Circuits*, 1(1):3.

Hunter, J. N. and Born, R. T. (2011). Stimulus-dependent modulation of suppressive influences in MT. *J Neurosci*, 31(2):678–86.

Isik, L., Leibo, J. Z., and Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Front Comp Neurosci*, 6:37.

Ito, M. and Gilbert, C. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22(3):593–604.

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

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell*, 20(11):1254–1259.

Jamalian, A. and Hamker, F. H. (2016). Biologically-Inspired Models for Attentive Robot Vision: A Review. In Pal, R., editor, *Innovative Research in Attention Modeling and Computer Vision Applications*, pages 69–98. IGI Global.

James, W. (1890). *The principles of psychology*. Henry Holt and Company.

Jannati, A., Spalek, T. M., and Di Lollo, V. (2013). A novel paradigm reveals the role of reentrant visual processes in object substitution masking. *Atten Percept Psychophys*, 75(6):1118–1127.

Jensen, O., Kaiser, J., and Lachaux, J.-P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends Neurosci*, 30(7):317–24.

Jones, H. E., Wang, W., and Sillito, a. M. (2002). Spatial organization and magnitude of orientation contrast interactions in primate V1. *J Neurophysiol*, 88(5):2796–808.

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.

Kastner, S., Pinsk, M. a., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22:751–761.

Kastner, S. and Ungerleider, S. (2000). Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci*, 23:315–341.

Kermani Kolankeh, A., Teichmann, M., and Hamker, F. H. (2015). Competition improves robustness against loss of information. *Front Comp Neurosci*, 9:35.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol*, 10(11):e1003915.

Kietzmann, T. C., Lange, S., and Riedmiller, M. (2009). Computational object recognition: a biologically motivated approach. *Biol Cybern*, 100(1):59–79.

Kiss, M., Velzen, J. V., and Eimer, M. (2008). The N2pc component and its links to attention shifts and spatially selective visual processing. *Psychophysiology*, 45(2):240–249.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol.*, 4(4):219–27.

Koechlin, E. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, 302(5648):1181–1185.

Krauzlis, R. J., Lovejoy, L. P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annu Rev Neurosci*, 36:165–182.

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.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *bioRxiv*, 029876.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2013). Imagenet classification with deep convolutional neural networks. In *Adv Neural Inf Process Syst 25 - NIPS 2012*, pages 1097–1105.

LaBerge, D. L. (1990). Attention. *Psychol Sci*, 1(3):156–162.

Lanyon, L. J. and Denham, S. L. (2004). A model of active visual search with object-based attention guiding scan paths. *Neural Netw*, 17:873–897.

Lanyon, L. J. and Denham, S. L. (2009). Modelling attention in individual cells leads to a system with realistic saccade behaviours. *Cognitive Neurodynamics*, 3:223–242.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. In *Proc of the IEEE*, volume 86, pages 2278–2324.

Lee, B.-T. and McPeek, R. M. (2013). The effects of distractors and spatial precues on covert visual search in macaque. *Vision Res*, 76:43–9.

Lee, J. and Maunsell, J. H. R. (2009). A normalization model of attentional modulation of single unit responses. *PLoS ONE*, 4(2):e4651.

Lee, J. and Maunsell, J. H. R. (2010a). Attentional modulation of MT neurons with single or multiple stimuli in their receptive fields. *J Neurosci*, 30(8):3058–66.

Lee, J. and Maunsell, J. H. R. (2010b). The effect of attention on neuronal responses to high and low contrast stimuli. *J Neurophysiol*, 104(2):960–71.

Lessmann, M. and Würtz, R. P. (2014). Learning invariant object recognition from temporal correlation in a hierarchical network. *Neural Netw*, 54:70–84.

Li, J., Tian, Y., Huang, T., and Gao, W. (2010). Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video. *Int J Comput Vis*, 90(2):150–165.

Li, X., Lu, Z.-L., Tjan, B. S., Dosher, B. a., and Chu, W. (2008). Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. *Proc Natl Acad Sci USA*, 105(16):6202–7.

Li, Y., Zhou, Y., Yan, J., Niu, Z., and Yang, J. (2009). Visual saliency based on conditional entropy. In *Proc Asian Conf Computer Vision 2009 - ACCV 2009*, pages 246–257.

Liang, J. and Qin, A. (2006). Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE Trans Evol Comput*, 10(3):281–295.

Liebold, B., Richter, R., Teichmann, M., Hamker, F. H., and Ohler, P. (2015). Human Capacities for Emotion Recognition and their Implications for Computer Vision. *I-Com*, 14(2):126–137.

Linde, O. and Lindeberg, T. (2012). Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition. *Comp Vis Image Und*, 116(4):538–560.

Lindsay, P., Taylor, M., and Forbes, S. (1968). Attention and multidimensional discrimination1. *Percept Psychophys*, 4(2):113–117.

Ling, S., Liu, T., and Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Res*, 49(10):1194–1204.

Lleras, A. and Moore, C. M. (2003). When the target becomes the mask: Using apparent motion to isolate the object-level component of object substitution masking. *J Exp Psychol Hum Percept Perform*, 29(1):106–120.

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annu Rev Psychol*, 55:207–34.

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

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, 60(2):91–110.

Luck, S. J. and Hillyard, S. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, 31(3):291–308.

Luiga, I. and Bachmann, T. (2008). Luminance processing in object substitution masking. *Vision Res*, 48(7):937–45.

Lyu, S. (2005). Mercer kernels for object recognition with local features. In *Proc IEEE Conf Computer Vision and Pattern Recognition 2005 - CVPR 2005*, pages 223–229.

MacEvoy, S., Tucker, T., and Fitzpatrick, D. (2009). A precise form of divisive suppression supports population coding in the primary visual cortex. *Nat Neurosci*, 12(5):637–645.

Martínez-Trujillo, J. and Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, 35(2):365–70.

Martinez-Trujillo, J. and Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr Bio*, 14:744–751.

Masland, R. H. (2001). The fundamental plan of the retina. *Nat Neurosci*, 4(9):877–886.

Maunsell, J. H. R. and Treue, S. (2006). Feature-based attention in visual cortex. *Trends Neurosci*, 29(6):317–22.

Mazza, V. and Pagano, S. (2014). Individual differences in masking sensitivity correlate with neural individuation abilities. In *Perception 43 ECVP Abstract Supplement*, page 99.

McAdams, C. J. and Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci*, 19(1):431–41.

Miau, F., Papageorgiou, C., and Itti, L. (2001). Neuromorphic algorithms for computer vision and attention. In Bosacchi, B., Fogel, D. B., and Bezdek, J. C., editors, *Proc Ann Int Symp Optical Science and Technology 2001 - ISOST 2001*, pages 12–23.

Miller, E., Gochin, P., and Gross, C. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Res*, 616:25–29.

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.

Mitchell, J., Stoner, G., and Reynolds, J. (2004). Object-based attention determines dominance in binocular rivalry. *Nature*, 429:410–413.

Mitchell, J. F., Stoner, G. R., Fallah, M., and Reynolds, J. H. (2003). Attentional selection of superimposed surfaces cannot be explained by modulation of the gain of color channels. *Vision Res*, 43(12):1323–1328.

Mitchell, J. F., Sundberg, K. A., and Reynolds, J. H. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron*, 55(1):131–41.

Mitri, S., Frintrop, S., Pervölz, K., and Surmann, H. (2005). Robust object detection at regions of interest with an application in ball recognition. In *Proc IEEE Int Conf Robotics and Automation 2005 - ICRA 2005*, pages 126–131.

Mobahi, H., Collobert, R., and Weston, J. (2009). Deep learning from temporal coherence in video. In *Proc 26th Ann Int Conf Machine Learning - ICML 2009*, pages 1–8.

Moore, C., Yantis, S., and Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. *Psychol Sci*, 9(2):104–110.

Moore, C. M. and Lleras, A. (2005). On the role of object representations in substitution masking. *J Exp Psychol Hum Percept Perform*, 31(6):1171–80.

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.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J Neurophysiol*, 70(3):909–19.

Motter, B. C. (2009). Central V4 receptive fields are scaled by the V1 cortical magnification and correspond to a constant-sized sampling of the V1 surface. *J Neurosci*, 29(18):5749–57.

Murray, S. (2008). The effects of spatial attention in early human visual cortex are stimulus independent. *J Vis*, 8(10):2.

Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision Res*, 45(2):205–31.

Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proc IEEE Conf Computer Vision and Pattern Recognition 2006 - CVPR 2006*, pages 2049–2056.

Neisser, U. (1967). *Cognitive Psychology*. Prentice-Hall.

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

Neumann, H. and Sepp, W. (1999). Recurrent V1-V2 interaction in early visual boundary processing. *Biol Cybern*, 81:425–444.

Ni, A. M., Ray, S., and Maunsell, J. H. R. (2012). Tuned normalization explains the size of attention modulations. *Neuron*, 73(4):803–13.

Ninomiya, T., Sawamura, H., Inoue, K.-I., and Takada, M. (2012). Segregated pathways carrying frontally derived top-down signals to visual areas MT and V4 in macaques. *J Neurosci*, 32(20):6851–8.

Nishida, S. (2011). Advancement of motion psychophysics: Review 2001-2010. *J Vis*, 11(5):11.

O'Craven, K., Downing, P., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401:584–587.

Oliva, A. and Torralba, A. (2003). Top-down control of visual attention in object detection. In *Proc IEEE Int Conf Image Processing 2003 - ICIP 2003*, pages 253–56.

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.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiol Rev*, 88:59–89.

O'Reilly, R. C. and Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput*, 18(2):283–328.

Ouerhani, N. (2003). *Visual Attention: From Bio-Inspired Modeling to Real-Time Implementation*. Doctoral dissertation, Institut de Microtechnique, Université de Neuchâtel, Lausanne, Switzerland.

Ozeki, H., Sadakane, O., Akasaki, T., Naito, T., Shimegi, S., and Sato, H. (2004). Relationship between excitation and inhibition underlying size tuning and contextual response modulation in the cat primary visual cortex. *J Neurosci*, 24(6):1428–38.

Põder, E. (2013). Attentional Gating Models of Object Substitution Masking. *J Exp Psychol Gen*, 142(4):1130–1141.

Põder, E. (2014). The changing picture of object substitution masking: reply to Di Lollo (2014). *Front Psychol*, 5(3):1004.

Pack, C. C., Hunter, J. N., and Born, R. T. (2005). Contrast dependence of suppressive influences in cortical area MT of alert macaque. *J Neurophysiol*, 93(3):1809–15.

Pagano, S. (2012). *Electrophysiological Correlates of Multiple Object Processing*. Doctoral dissertation, Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy.

Paletta, L., Fritz, G., and Seifert, C. (2005). Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proc 22nd Int Conf Machine Learning - ICML 2005*, pages 649–656.

Palomino, A. J., Marfil, R., Bandera, J. P., and Bandera, A. (2011). A Novel Biologically Inspired Attention Mechanism for a Social Robot. *EURASIP J Adv Sig Pr*, 2011:1–10.

Pang, D., Kimura, A., Takeuchi, T., Yamato, J., and Kashino, K. (2008). A stochastic model of human visual attention with a dynamic Bayesian network. In *Proc IEEE Int Conf Multimedia & Expo 2008 - ICME 2008*, pages 1073 – 1076.

Parker, A. and Hawken, M. (1988). Two-dimensional spatial structure of receptive fields in monkey striate cortex. *J Opt Soc Am A*, 5:598.

Parks, D. H. and Levine, M. D. (2006). McGill Object Detection Suite. In *Proc 3rd Canadian Conf Computer and Robot Vision 2006 - CRV 2006*, page 47.

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

Pessoa, L. and Adolphs, R. (2010). Emotion processing and the amygdala: from a'low road'to'many roads' of evaluating biological significance. *Nat Rev Neurosci*, 11(11):773–783.

Peters, R. J. and Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proc IEEE Conf Computer Vision and Pattern Recognition 2007 - CVPR 2007*, number 1, pages 1–8.

Petersen, S. and Posner, M. (2012). The attention system of the human brain: 20 years after. *Annu Rev Neurosci*, pages 73–89.

Pilling, M. and Gellatly, A. (2010). Object substitution masking and the object updating hypothesis. *Psychon Bull Rev*, 17(5):737–42.

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput Biol*, 4(1):e27.

Poort, J., Raudies, F., Wannig, A., Lamme, V. a. F., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, 75(1):143–56.

Posner, M. I. (1980). Orienting of attention. *J Exp Psychol*, 32:3–25.

Posner, M. I. and Petersen, S. E. (1990). The attention system of the human brain. *Annu Rev Neurosci*, 13:25–42.

Posner, M. I. and Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annu Rev Psychol*, 58:1–23.

Potjans, T. C. and Diesmann, M. (2012). The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. *Cerebral cortex*, 24(3):785–806.

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.

Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychol Bull*, 124(2):123–64.

Prime, D. J., Pluchino, P., Eimer, M., Dell'Acqua, R., and Jolicœ ur, P. (2011). Object-substitution masking modulates spatial attention deployment and the encoding of information in visual short-term memory: insights from occipito-parietal ERP components. *Psychophysiology*, 48(5):687–96.

Qian, N. (1994). Computing Stereo Disparity and Motion with Known Binocular Cell Properties. *Neural Comput*, 6(3):390–404.

Qiu, F. T., Sugihara, T., and von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nat Neurosci*, 10(11):1492–99.

Rasolzadeh, B., Björkman, M., Huebner, K., and Kragic, D. (2009). An Active Vision System for Detecting, Fixating and Manipulating Objects in the Real World. *Int J Robot Res*, 29(2-3):133–154.

Ratliff, F., Knight, B. W., Toyoda, J.-i., and Hartline, H. K. (1967). Enhancement of Flicker by Lateral Inhibition. *Science*, 158(3799):392–393.

Raudies, F., Mingolla, E., and Neumann, H. (2011). A Model of Motion Transparency Processing with Local Center-Surround Interactions and Feedback. *Neural Comput*, 23:2868–2914.

Raudies, F. and Neumann, H. (2009). An efficient linear method for the estimation of ego-motion from optical flow. In Denzler, J., Notni, G., and Herbert, S., editors, *Pattern Recognition, Lecture Notes in Computer Science 5748*, pages 11–20.

Raudies, F. and Neumann, H. (2010). A neural model of the temporal dynamics of figure-ground segregation in motion perception. *Neural Netw*, 23:160–176.

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.

Read, J. C. a. and Cumming, B. G. (2006). Does depth perception require vertical-disparity detectors? *J Vis*, 6(12):1.

Reeves, A. and Sperling, G. (1986). Attention gating in short-term visual memory. *Psychol Rev*, 93(2):180–206.

Rensink, R. a. (2002). Change detection. *Annu Rev Psychol*, 53:245–77.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci*, 19(5):1736–53.

Reynolds, J. H. and Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, 37(5):853–63.

Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61(2):168–85.

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–14.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2(11):1019–25.

Ritter, M. (2013). *Optimierung von Algorithmen zur Videoanalyse — Ein Analyseframework fur die Anforderungen lokaler Fernsehsender*. Doctoral dissertation, Technische Universität Chemnitz, Germany.

Robitaille, N. and Jolicoeur, P. (2006). Fundamental properties of the N2pc as an index of spatial attention: Effects of masking. *Can J Exp Psychol*, 60(2):79–89.

Rodríguez-Sánchez, A. J., Simine, E., and Tsotsos, J. K. (2007). Attention and visual search. *Int J Neural Syst*, 17(4):275–288.

Roelfsema, P., Lamme, V., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395:376–381.

Rolls, E. T. (2012). Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet. *Front Comp Neurosci*, 6:35.

Rothenstein, A. L., Rodríguez-Sánchez, A. J., Simine, E., and Tsotsos, J. K. (2008). Visual Feature Binding Within the Selective Tuning Attention Framework. *Int J Pattern Recogn Artif Intell*, 22(5):861–881.

Rothenstein, A. L. and Tsotsos, J. K. (2014). Attentional modulation and selection - An integrated approach. *PLoS ONE*, 9(6):e99681.

Russell, A. F., Mihalaş, S., von der Heydt, R., Niebur, E., and Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Res*, 94:1–15.

Rybak, I. a., Gusakova, V. I., Golovan, a. V., Podladchikova, L. N., and Shevtsova, N. a. (1998). A model of attention-guided visual perception and recognition. *Vision Res*, 38(15-16):2387–400.

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.

Sakai, K. (2008). Task set and prefrontal cortex. *Annu Rev Neurosci*, 31:219–45.

Salah, A., Alpaydin, E., and Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans Pattern Anal Mach Intell*, 24(3):420–425.

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.

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Trans Pattern Anal Mach Intell*, 5:530–534.

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., and Leventhal, a. G. (1998). Signal timing across the macaque visual system. *J Neurophysiol*, 79(6):3272–8.

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.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). Over-Feat : Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc Int Conf Learning Representations 2014 - ICLR 2014*, pages 1–15.

Serre, T. (2006). *Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines*. Doctoral dissertation, Massachusetts Institute of Technology, USA.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). A Theory of Object Recognition : Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. Technical report.

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

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Trans Pattern Anal Mach Intell*, 29(3):411.

Shapiro, K., Raymond, J., and Arnell, K. (1997). The attentional blink. *Trends Cogn Sci*, 1(8):291–296.

Shelley-Tremblay, J. and Mack, A. (1999). Metacontrast Masking and Attention. *Psychol Sci*, 10(6):508–515.

Siegel, R. and Read, H. (1997). Analysis of optic flow in the monkey parietal area 7a. *Cerebral Cortex*, 7(4):327–346.

Sigala, N., Gabbiani, F., and Logothetis, N. K. (2002). Visual categorization and object representation in monkeys and humans. *J Cogn Neurosci*, 14(2):187–98.

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.

Smith, P. L. and Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychol Rev*, 116(2):283–317.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf Process Manage*, 45(4):427–437.

Sperling, G. (1960). The information available in brief visual presentations. *Psychol Monogr: Gen Appl*, 74(11):1–29.

Sperling, G. and Weichselgartner, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychol Rev*, 102:503–532.

Sprague, N. and Ballard, D. (2004). Eye movements for reward maximization. In *Adv Neural Inf Process Syst 16 - NIPS 2003*, pages 1467–1474.

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

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Res*, 48(12):1391–408.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *J Neurosci*, 30(9):3531–43.

Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Comput*, 24(1):60–103.

Spratling, M. W. and Johnson, M. H. (2004). A feedback model of visual attention. *J Cogn Neurosci*, 16(2):219–37.

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.

Steenrod, S. C., Phillips, M. H., and Goldberg, M. E. (2013). The lateral intraparietal area codes the location of saccade targets and not the dimension of the saccades that will be made to acquire them. *J Neurophysiol*, 109(10):2596–605.

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.

Sun, Y. and Fisher, R. (2003). Object-based visual attention for computer vision. *Artif Intelli*, 146(1):77–123.

Sundberg, K. A., Mitchell, J. F., Gawne, T. J., and Reynolds, J. H. (2012). Attention influences single unit and local field potential response latencies in visual cortical area V4. *J Neurosci*, 32(45):16040–50.

Sundberg, K. A., Mitchell, J. F., and Reynolds, J. H. (2009). Spatial attention modulates center-surround interactions in macaque visual area v4. *Neuron*, 61(6):952–63.

Swets, J. A. (2014). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Psychology Press.

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

Tata, M. and Giaschi, D. (2004). Warning: Attending to a mask may be hazardous to your perception. *Psychon Bull Rev*, 11(2):262–268.

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.

Thielscher, A., Kölle, M., Neumann, H., Spitzer, M., and Grön, G. (2008). Texture segmentation in human perception: a combined modeling and fMRI study. *Neuroscience*, 151(3):730–6.

Thielscher, A. and Neumann, H. (2005). Neural mechanisms of human texture processing: texture boundary detection and visual search. *Spatial Vision*, 18(2):227–257.

Thielscher, A. and Neumann, H. (2007). A computational model to link psychophysics and cortical cell activation patterns in human texture processing. *J Comp Neurosci*, 22:255–282.

Thielscher, A. and Neumann, H. (2008). Globally consistent depth sorting of overlapping 2D surfaces in a model using local recurrent interactions. *Biol Cybern*, 98:305–337.

Thomson, A., West, D., Wang, Y., and Bannister, A. (2002). Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2–5 of adult rat and cat neocortex: triple intracellular recordings and. *Cerebral cortex*, 12:936–953.

Todd, J. J., Fougnie, D., and Marois, R. (2005). Visual short-term memory load suppresses temporo-parietal junction activity and induces inattentional blindness. *Psychol Sci*, 16(12):965–72.

Todd, J. J. and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428:751–54.

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.

Tootell, R. B. H., Mendola, J. D., Hadjikhani, N. K., Ledden, P. J., Liu, A. K., Reppas, J. B., Sereno, M. I., and Dale, A. M. (1997). Functional Analysis of V3A and Related Areas in Human Visual Cortex. *J Neurosci*, 17(18):7060–7078.

Torralba, A. (2003a). Contextual priming for object detection. *Int J Comput Vis*, 53(2):169–191.

Torralba, A. (2003b). Modeling global scene factors in attention. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1407–18.

Treisman, A. (1964). Selective Attention in Man. *Br Med Bull*, 20(1):12–16.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychol*, 136:97–136.

Treue, S. and Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, 382:539–41.

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

Tsotsos, J. K., Culhane, S., Wai, W. K., and Lai, Y. (1995). Modeling visual attention via selective tuning. *Artif Intell*, 78:507–545.

Tsotsos, J. K., Liu, Y., Martinez-Trujillo, J. C., Pomplun, M., Simine, E., and Zhou, K. (2005). Attending to visual motion. *Comput Vis Image Underst*, 100:3–40.

Tsotsos, J. K., Rodríguez-Sánchez, A. J., Rothenstein, A. L., and Simine, E. (2008). The different stages of visual recognition need different attentional binding strategies. *Brain Res*, 1225:119–32.

Tsotsos, J. K. and Rothenstein, A. L. (2011). Computational models of visual attention. *Scholarpedia*, 6(1):6201.

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

Ungerleider, L. and Haxby, J. (1994). 'What'and 'where'in the human brain. *Curr Opin Neurobiol*, 4:157–165.

Ungerleider, L. G., Galkin, T. W., Desimone, R., and Gattass, R. (2008). Cortical connections of area V4 in the macaque. *Cerebral Cortex*, 18(3):477–99.

Vincent, B. T., Troscianko, T., and Gilchrist, I. D. (2007). Investigating a space-variant weighted salience account of visual selection. *Vision Res*, 47(13):1809–20.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *Int J Comput Vis*, 57(2):137–154.

Viswanathan, L. and Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. *Perception*, 31:1415–1437.

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.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Voss, Leipzig.

Voorhees, H. and Poggio, T. (1988). Computing texture boundaries from images. *Nature*, 333:364–367.

Wagatsuma, N., Potjans, T. C., Diesmann, M., Sakai, K., and Fukai, T. (2013). Spatial and feature-based attention in a layered cortical microcircuit model. *PLoS ONE*, 8(12):e80788.

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

Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw*, 19(9):1395–407.

Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2004). On the usefulness of attention for object recognition. In *Proc Workshop Attention and Performance in Computational Vision at ECCV*, pages 96–103.

Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Comp Vis Image Und*, 100(1-2):41–63.

Walther, D. B. and Koch, C. (2007). Attention in hierarchical models of object recognition. *Prog Brain Res*, 165:57–78.

Ward, L. M. (2008). Attention. 3(10):1538.

Wässle, H. (2004). Parallel processing in the mammalian retina. *Nat Revs Neurosci*, 5(10):747–757.

Wässle, H. and Boycott, B. (1991). Functional architecture of the mammalian retina. *Physiol Rev*, 71(2):447–80.

Weidenbacher, U. and Neumann, H. (2009). Extraction of surface-related features in a recurrent model of V1-V2 interactions. *PLoS ONE*, 4(6):e5909.

Weisstein, N. (1968). A Rashevsky-Landahl neural net: Simulation of metacontrast. *Psychol Rev*, 75:494–521.

Westphal, G. and Würtz, R. (2009). Combining feature-and correspondence-based methods for visual object recognition. *Neural Comput*, 21(7):1952–1989.

Whitney, D. and Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends Cogn Sci*, 15(4):160–8.

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.

Williford, T. and Maunsell, J. H. R. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *J Neurophysiol*, 96(1):40–54.

Willmore, B. D. B., Prenger, R. J., and Gallant, J. L. (2010). Neural representation of natural images in visual area V2. *J Neurosci*, 30(6):2102–14.

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.

Wischnewski, M., Belardinelli, A., Schneider, W. X., and Steil, J. J. (2010). Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. *Cog Comp*, 2(4):326–343.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychon Bull Rev*, 1(2):202–238.

Wolfe, J. M. (2000). The deployment of visual attention: Two surprises. In *Search and Target Acquisition*, pages 2011–2021.

Woodman, G. and Luck, S. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature*, 400:867–869.

Woodman, G. F. (2010). Masked targets trigger event-related potentials indexing shifts of attention but not error detection. *Psychophysiology*, 47(3):410–4.

Woodman, G. F. and Luck, S. J. (2003). Dissociations among Attention, Perception, and Awareness during Object-Substitution Masking. *Psychol Sci*, 14(6):605–11.

Wundt, W. M. (1874). *Grundzüge der physiologischen Psychologie*. Wilhelm Engelmann, Leipzig.

Xu, T. and Chenkov, N. (2009). Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots. In *Proc IEEE/RSJ Conf Intelligent Robots and Systems 2009 - IROS2009*, pages 4009–4014.

Yanulevskaya, V., Uijlings, J., Geusebroek, J.-m., Sebe, N., and Smeulders, A. (2013). A proto-object-based computational model for visual saliency. *J Vis*, 13(13):27.

Yu, A., Mann, G. K. I., and Gosine, R. G. (2008). An object-based visual attention model for robots. In *Proc IEEE Int Conf Robotics and Automation 2008 - ICRA 2008*, pages 943–948.

Yu, Y., Gu, J., Mann, G. K. I., and Gosine, R. G. (2013). Development and Evaluation of Object-Based Visual Attention for Automatic Perception of Robots. *IEEE Trans Autom Sci Eng*, 10(2):365–379.

Yu, Y., Mann, G. K. I., and Gosine, R. G. (2010). An object-based visual attention model for robotic applications. *IEEE Trans Syst. Man Cybern: Cybern*, 40(5):1398–412.

Zaharescu, A. (2004). *Towards a Biologically Plausible Active Visual Search Model*. Doctoral dissertation, York University, Toronto, Canada.

Zanos, T. P., Mineault, P. J., Monteon, J. a., and Pack, C. C. (2011). Functional connectivity during surround suppression in macaque area V4. In *Proc. Ann. Int. Conf. Engineering in Medicine and Biology Society 2011 - EMBS 2011*, volume 2011, pages 3342–5.

Zeiler, M. D. (2012). *Hierarchical Convolutional Deep Learning in Computer Vision*. Doctoral dissertation, New York University (NYU), USA.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proc European Conf Computer Vision 2014 - ECCV 2014*, pages 1–11.

Zhou, H. and Desimone, R. (2011). Feature-Based Attention in the Frontal Eye Field and Area V4 during Visual Search. *Neuron*, 70(6):1205–17.

Ziesche, A. and Hamker, F. H. (2011). A computational model for the influence of corollary discharge and proprioception on the perisaccadic mislocalization of briefly presented stimuli in complete darkness. *J Neurosci*, 31(48):17392–17405.

Ziesche, A. and Hamker, F. H. (2014). Brain circuits underlying visual stability across eye movements-converging evidence for a neuro-computational model of area LIP. *Front Comp Neurosci*, 8:25.

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

Zirnsak, M., Gerhards, R. G. K., Kiani, R., Lappe, M., and Hamker, F. H. (2011b). Anticipatory Saccade Target Processing and the Presaccadic Transfer of Visual Features. *J Neurosci*, 31(49):17887–17891.

Zirnsak, M. and Hamker, F. H. (2010). Attention Alters Feature Space in Motion Processing. *J Neurosci*, 30(20):6882–6890.

Zirnsak, M., Lappe, M., and Hamker, F. H. (2010). The spatial distribution of receptive field changes in a model of peri-saccadic perception: predictive remapping and shifts towards the saccade target. *Vision Res*, 50(14):1328–1337.

Zirnsak, M. and Moore, T. (2014). Saccades and shifting receptive fields: anticipating consequences or selecting targets? *Trends Cogn Sci*, 18(12):621–628.

Zirnsak, M., Steinmetz, N. a., Noudoost, B., Xu, K. Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. *Nature*, 507:504–507.