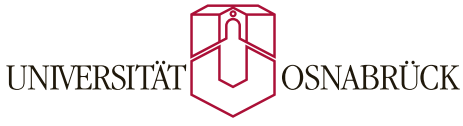


Computational and neural models of oculomotor control.

Dissertation
zur Erlangung des Grades
"Doktor der Naturwissenschaft"
im Fachbereich Humanwissenschaften
der Universität Osnabrück

vorgelegt von
Niklas Wilming

Osnabrück, im September 2014

Supervisor:

Prof. Dr. Peter König
University of Osnabrück, Osnabrück, Germany

Prof. Dr. Elizabeth Buffalo
Washington University, Seattle, USA

Additional reviewer:

Prof. Dr. Frank Jäkel
University of Osnabrück, Osnabrück, Germany

Prof. Dr. Gordon Pipa
University of Osnabrück, Osnabrück, Germany

Curriculum Vitae

Niklas Wilming

Kiwittstr. 19
49080 Osnabrück
Germany

nwilming@uos.de
+49 179 921 0065

Journal Articles[1]

2014	*Wilming, N., Jutras, M., Buffalo, E. A., & König, P. (n.d.). Differential contribution of low and high-level image content to eye movements in monkeys and humans. *(In Preparation)*
2013	*Betz, T., Wilming, N., Bogler, C., Haynes, J.-D., & König, P. (2013). Dissociation between saliency signals and activity in early visual cortex. *Journal of Vision*, *13*(14), 1–12
2013	*Wilming, N., Harst, S., Schmidt, N., & König, P. (2013, January). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Computational Biology*, *9*(1), e1002871
2013	*König, P., Wilming, N., Kaspar, K., Nagel, S. K., & Onat, S. (2013). Predictions in the light of your own action repertoire as a general computational principle. *Behavioral and Brain Sciences*, *36*(03), 39–40

[1] Articles marked with an asterisk are part of this thesis.

2011 | *Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and Limits of Models of Fixation Selection. *PLoS ONE*, *6*(9), e24038

2010 | Betz, T., Kietzmann, T. C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, *10*, 1–14

Conference Contributions[2]

2012 | Wilming, N., Harst, S., Schmidt, N., König, P. (2012). Facilitation and Speed-up of Return Saccades Contribute to an Optimal Foraging strategy. In *"Comptetition and Priority Control in Mind And Brain: New Perspectives from Task-Driven Vision"*

2012 | + Kootstra, G., Wilming, N., Schmidt, N., Djurfeldt, M., Kragic, D., & Peter, K. (2012). Learning and Adaptation of Sensorimotor Contingencies : Prism-Adaptation, a case study. In *From animals to animats 12 - 12th international conference on simulation of adaptive behavior proceedings.* (Vol. 7426/2012, pp. 341–350). Springer

2012 | Wilming, N., Harst, S., Schmidt, N., König, P. (2012) Facilitation of return fixations but delay of turning; revising the concept of IOR. *5th International Conference on Cognitive Systems.*

2011 | Wilming, N., Betz, T., Harst, S., Waterkamp, S., König, P. (2011). Bayesian Modelling of Eye-Movements based on an Analysis of the Conditional Dependence Structure between consecutive Saccades. *Osnabrück Computational Cognition Alliance Meeting*.

[2] + marks peer-reviewed contributions.

2011 Wilming, N., Betz, T., Kietzmann, TC., König, P. (2011). How well can we model human overt attention? *Interdisciplinary College 2011*

2010 Wilming, N., Betz, T., Schreiber, C., and König, P. (2010). Capabilities and limitations of bottom-up salience modelling of visual attention. *FENS Abstr. Vol 5.*

2009 + Wilming, N., Wolfsteller, F., König, P., Caseiro, R., Xavier, J., & Araújo, H. (2009). Attention Models for Vergence Movements based on the JAMF Framework and the POPEYE Robot. In *Isapp 2009 - proceedings of the fourth international conference on computer vision theory and applications, lisboa, portugal, february 5-8, 2009 - volume 2.* (pp. 429–437). INSTICC Press

2009 Schreiber, C., Betz, T., Wilming, N., Kietzmann, TC., and König, P. (2009). Task-effects on Viewing Behavior Examined in School Children. *8th Göttingen Meeting of the German Neuroscience Society*.

2009 + Steger, J., Wilming, N., Wolfsteller, F., Höning, N., & König, P. (2009). The JAMF Attention Modelling Framework. In *Attention in cognitive systems. lecture notes in computer science.* (pp. 153–165). Berlin Heidelberg: Springer

Talks

2012 Do actions change our perception? *Donders Discussion, Nijmegen, the Netherlands*

2011 Human overt attention under natural conditions. *KTH Royal Institute of Technology in Stockholm*

2011 Lower and upper bounds for the predictability of viewing behavior. *Short talk at the Osnabrück Computational Cognition Alliance Meeting on "Natural Computation in Hierarchies"*

Supervised B.Sc. and M.Sc. theses

2014	The influence of power and precision grip on overt visual attention. B.Sc. thesis of Tabea Kossen
2014	The effect of head movements on viewing behaviour. B.Sc. thesis of Alexandra Vormberg
2013	Is Information from previous fixations incorporated into a prior of bayesian-optimal search behavior? M.Sc. thesis of Simon Harst.
2012	Prism adaptation in human pointing: Recalibration and realignment aftereffects for distal and proximal pointing to targets of varying degrees. B.Sc. thesis of Ulrike Kuhl.
2011	Higher order properties of gaze movement trajectories and inhibition of return. B.Sc. thesis of Simon Harst.
2011	Investigating top-down and bottom-up influences on overt visual attention in autism spectrum disorder. M.Sc. thesis of Hannah Knepper.
2010	An adaptive model of human attention based on hierarchical feature complexity. M.Sc. thesis of Anna Metzger.

Teaching

SS2014	Action & Cognition II
WS2013	Occulomotor control and attention
SS2011	Scientific computing with Python
WS2010	Probability Theory: From Extended Logic to Applications in Cognitive Science

Abstract

Seeing is more than sight: it is the entire action-perception loop involved in taking in the world around us. Unlike a camera, our eyes can only resolve a small part of the environment sharply. Therefore, we must constantly move our eyes to scrutinise the parts of our environment that seem most worthy of our highest visual acuity. Eye movements are thus the observable consequences of a complex and crucial decision-making process that is fundamental to how we interact with the world.

This thesis investigates properties and the neural basis of eye-movement behavior in humans and monkeys. In the interdisciplinary tradition of cognitive science, the thesis spans fields and utilizes computational models as explanatory vehicles. A central theme is the so-called saliency map model of attention, the de facto computational model of viewing behavior.

The saliency map model assumes that attention is directed at the peaks of a map that encodes the saliency of locations in the visual field. Saliency can roughly be thought of as how worthy a location is of attention. It forms a common currency that allows different processes to influence the distribution of attention.

The four different studies in this thesis provide four different perspectives on viewing behavior and the saliency map model. The first study establishes a methodology to evaluate the predictive power of models of viewing behavior, and determines which properties of viewing behavior are important for this evaluation. Applying this

methodological foundation to the saliency map model reveals that state-of-the-art models do not provide satisfactory explanations of viewing behavior. The second study investigates spatio-temporal properties of eye-movements, finding that observers often re-fixate locations in pictures and that their eye movements possess a rich spatio-temporal structure. These results speak directly against a causal role of "inhibition of return", which is a popular component of many saliency map models. The third study shifts focus to the neural basis of the oculomotor behaviour. fMRI is used to probe the relationship between the computation of saliency and actual processing in the brain. Our results, in contrast to those of other studies, suggest that early visual areas do not compute saliency, but instead compute visual features upon which the saliency map operates. Much of what we know about the neural basis of oculomotor control comes from invasive studies in animals, but it is unclear to what extent saliency computations are comparable between species. Thus, the fourth study compares the viewing behavior of monkeys and humans, to look for evidence of the same underlying processes. We find a strong similarity between the species in saliency-driven viewing behavior. The many saliency-processing areas that have been identified in monkeys therefore likely have a role in saliency processing in the human brain as well.

This thesis contributes to our understanding of oculomotor control on multiple levels. The results in this thesis suggest that models of viewing behavior should treat saccade-target selection as a dynamic process where past decisions influence future decisions and where saliency varies over time. This selection process likely takes place in a distributed network in the brain which receives bottom-up input from early visual areas. Encouraged by these results, we speculate that normative and embodied models of cognition offer an explanation of oculomotor control that takes these results into account. In turn, explaining oculomotor control is an important part of the much deeper question of how our mind interacts with the world.

Contents

Glossary

AUC area under the curve.

BOLD blood oxygen level dependent.

EEG electroencephalography.

EPI echo planar image.

FDM fixation density map.

FEF frontal eye field.

fMRI functional magnetic resonance imaging.

FOV field of view.

FPR false positive rate.

FWHM full width half maximum.

GLM general linear model.

HRF hemodynamic response function.

IOR inhibition of return.

IT inferior temporal cortex.

LGN lateral geniculate nucleus.

LIP lateral intraparietal sulcus.

MST medial superior temporal area.

MT medial temporal area.

MVPA multivariate pattern analysis.

NSS normalized scanpath salience.

PET positron emission tomography.

RF receptive field.

RMSE root mean square error.

ROI region of interest.

SC superior colliculus.

V1 primary visual cortex.

V2 visual area 2.

V3 visual area 3.

V4 visual area 4.

WTA winner takes all.

Chapter 1

Making contact

1.1 Eye movements establish contact with the world.

The majority of our eye movements are automatic and, like our heartbeat, go unnoticed most of the time. A simple method to observe one's own eye movements is to attend to the movement of an after-image created by staring at a bright spot for some time (Yarbus (1967) mentions several other methods). After-images are the consequence of retinal adaptation to bright stimuli and the after-image itself is therefore stationary on the retina. Movement of the after-image relative to the environment must therefore be due to movement of the eyes. This technique allows the distinguishing of three different phases of eye movements: fixations, saccades and smooth pursuit. Fixations are periods where the eye is stable and shows very little movement. Saccades are fast ballistic movements between fixations, and smooth pursuit movements are slow eye movements when the eyes follow a moving object in the environment (see Figure 1.1).

What is the importance of these eye movements for the act of seeing? Because eye movements are difficult to observe introspectively, it is fruitful to explore the role of eye movements with an analogy where the relation between perception and action is more obvious. That is to say, we can learn a lot about visual perception when we

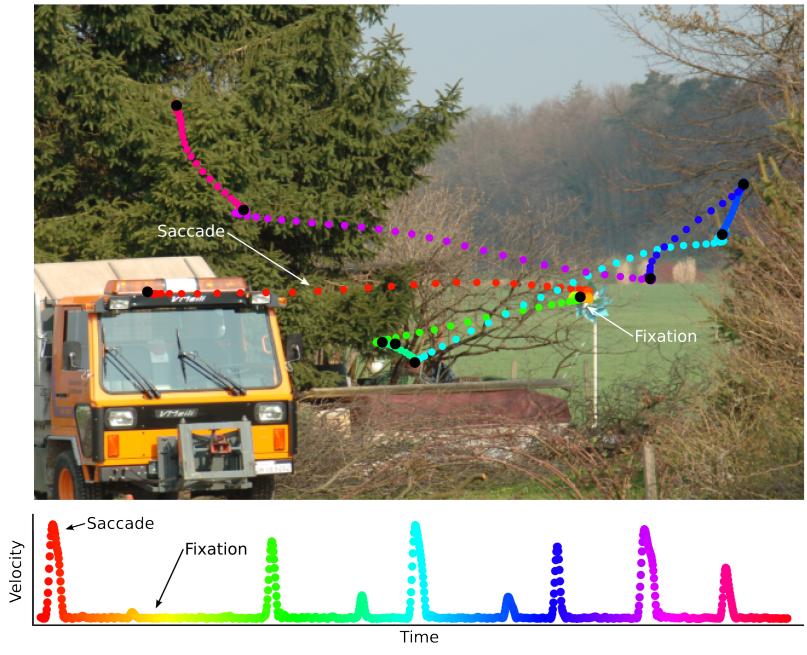


Figure 1.1: Fixations and saccades. Colored dots show the eye-position of one observer over the time course of ca. 3s. Eye-position is sampled with 500Hz and each dot shows one sample. Accordingly, dots are spaced far apart during saccades and closeby during fixations. Black dots show average gaze position during a fixation. The bottom panel shows the velocity of the eye at each sample location.

close our eyes and instead focus on how we experience our world through touch (see MacKay, 1973; and O'Regan, 1992, for similar analogies).

For example let's focus on the sensations in our hand when we place it on a flat surface. As soon as we establish contact, various neuronal receptors react to the physical changes of our hand. Immediately we can feel if the surface is cold or warm and bumpy or smooth. Our hand establishes contact between our mind and the outside world. To us, the physical changes in our hand do not ap-

pear as physical changes of the hand but as properties of the outside world. When we establish contact, we don't feel our body but instead feel the outside world[1].

Establishing contact by touch requires that we actively explore the world by moving our hands. They are simply too small to cover many objects. This is not always a limitation, as moving our hands allows us to feel properties of the world that are not directly transmitted by receptors in our hands. Fine textures, for example, are translated to sensations by vibrations of the finger that is moved over a texture (Scheibert, Leurent, Prevost, & Debregeas, 2009). When the finger is stopped, the sensation of the texture also stops. Some animals are experts in exploring the world by interpreting the spatiotemporal deflections of their whiskers (Diamond, von Heimendahl, Knutsen, Kleinfeld, & Ahissar, 2008). These are two cases where the combination of sensory input and active movement of a sensor allows the perception of properties of the world that are not contained in static sensory input.

A further important property of touch is that it brings objects into focus in their entirety, even though we can rarely touch entire objects at once. Consider for example the gaps between our fingers (O'Regan, 1992). These provide no information about objects that we touch, but objects do not feel like they have gaps. If you grasp for keys in your pocket, the keys feel like your keys, even if not all parts of the keys are being touched. Recognizing the keys brings the keys into focus and the actual sensation of touching the keys is secondary. This makes clear that our mind does not create a veridical "representation" of the sensory input in our mind. Instead, touching the world gives us direct access to properties (e.g. objects) of the world that are relevant for our behavior. Touch establishes contact between the mind and the physical world.

These examples illustrate that we can get to know the world by actively sampling the environment with our hands. And by doing so we situate ourselves in the world.

[1] Of course we can also focus on our hand and feel the changes of our hand.

Eye movements have similar properties. Like touch, sight brings parts of the world into focus, literally and figuratively. When we see, we see the world, we don't feel the electrical processes in our retina. But the analogy to touch is richer than this because, like touch, seeing is an active process. This can be illustrated with several examples.

First, the distribution of photo-receptors on our retinae is strongly heterogeneous. The highest density of receptors is found in the fovea, the area of highest visual acuity. The fovea itself is ca. 2° of visual angle in diameter (Curcio, Sloan, Packer, Hendrickson, & Kalina, 1987), which roughly corresponds to the width of the thumb held at arm's length (O'Shea, 1991). The consequences of uneven photo-receptor placement can be observed by closing one eye, fixating a letter in this text and then observing the sharpness of text that is further and further away from the fixated letter. This demonstrates that the physiological structure of our eyes forces us to constantly move our gaze location (by moving our eyes, head or body) to keep parts of the environment in focus.

A prime example for this behavior is reading. While reading, our eyes jump between words, sometimes even between letters in the same word (Rayner, 2009). Since we are effectively blind during a saccade (Bridgeman, Hendry, & Stark, 1975)[2], where we fixate dictates what we can read. In general, where we fixate determines what we perceive with highest acuity.

The importance of selecting appropriate fixation locations is reflected in specific and stereotyped eye movement patterns during everyday activities. Cricket[3] players, for example, carry out anticipatory saccades to the bounce of the ball (Land & McLeod, 2000). Similar anticipatory saccades are found during table tennis (Land & Furneaux, 1997), squash (Hayhoe, McKinney, Chajka, & Pelz, 2012) and baseball (Bahill & LaRitz, 1984). Eye movements during such activities, as well as others like tea-making (Land, Mennie, & Rusted, 1999), sandwich making (Hayhoe, 2000), copying (Ballard & Hayhoe,

[2]Which explains why we can't observe our own eye movements in a mirror.

[3]I admit, cricket is not an everyday activity for most humans.

1992), and driving (Sullivan, Johnson, Rothkopf, Ballard, & Hayhoe, 2012) are typically highly structured around the locus of action and anticipate the action. This is the case even during observation of others' actions (Flanagan & Johansson, 2003). Seeing therefore involves skillful use of eye movements.

Just how important the selection of fixation locations is for our perception is demonstrated in an extreme way by inattentional and change blindness paradigms. Simons and Chabris (1999) asked observers to watch a video of students passing around a ball and to count how often the ball was passed. After some time, an actor dressed up as a Gorilla walked through the scene. Surprisingly, many observers failed to notice the Gorilla. At the same time, they were confident that they would have noticed a Gorilla. Similarly, observers often fail to notice large changes in an image, such as removal of a lamp post from a street scene, when the two versions of the scene are flicked back and forth with a short period of black in between (O'Regan, Rensink, & Clark, 1999). These studies highlight that seeing likely does not consist of building an internal representation of the external world. Instead, some researchers have suggested that we use eye movements to quickly gain access to the world, instead of placing details of the world in working memory (O'Regan, 1992; O'Regan & Nöe, 2001). This stresses again that the act of seeing is an active process where sensory input and eye movements interact in a loop.

In this sensorimotor loop, the sensory evidence can influence the next motor actions and the motor actions can change what is perceived. The first dependence (sensory $\rightarrow$ motor), is demonstrated by the specificity of eye movement trajectories in tasks like squash (see above). Here, visual events trigger specific motor responses in the form of eye movements. The second dependence (motor $\rightarrow$ sensory) is exemplified by studies which show that, like in touch, seeing can exploit the spatiotemporal structure of the sensorimotor information to guide perception. Hafed and Krauzlis (2006) for example show that eye movements constrain the interpretation of

visual input. When observers saw an occluded chevron that moved on a circular trajectory, they had trouble recognizing that the visual input was caused by the chevron. In a movement condition that produced identical retinal stimulation, but where observers 'caused' the movement by moving their eyes on a circular path, observers had less trouble recognizing the chevron. Another example is given by Kuang, Poletti, Victor, and Rucci (2012), who demonstrate that the spatio-temporal structure of miniature eye movements effectively removes correlations in the visual input. This facilitates subsequent redundancy reduction and edge detection processes.

Another case in point is the structure of eye movements during the viewing of ambiguous stimuli (Kietzmann, Geuter, & König, 2011). Different interpretations of an ambiguous figure are preceded by eye movement trajectories that are specific to the subsequent interpretation. Seeing is therefore an active process that is critically dependent on eye movements.

In summary, I argue that eye movements are a key component in how we perceive and interact with our environment. Vision gives us the ability to connect our minds to the outside world and vision itself depends in a fundamental way on eye movements. This is reflected in the active nature of seeing, which is itself rendered evident by the specificity of task-driven eye movements and the ubiquity of eye movements.

The work presented in this thesis aims at furthering our understanding of how we decide where to direct our gaze. The hope is that understanding why, when and how we shift our eyes will contribute to an understanding of vision, and thereby to an understanding of how our minds can situate themselves within the physical world.

The manuscripts presented in this thesis revolve around a specific model of overt and covert attention shifts (the "saliency map" model by Koch and Ullman (1985) and Itti and Koch (2001a); a detailled explanation follows). While the manuscripts can be understood separately, my goal for this thesis is to relate the different manuscripts

to the saliency map model[4] and to make obvious how they advance our understanding of fixation selection processes. To this end, I will first explain how a model of covert and overt attention relates to eye movements. I will also review other approaches to understanding eye movements to give the reader a notion of what is not addressed in this thesis. The following sections then introduce the saliency map model, review its impact on the scientific community, and review physiological evidence for the model. After these sections have set the stage, I will explain how the different manuscripts relate to the saliency map model of covert and overt attention shifts.

1.2 Fixation selection and attention

What is the relation between eye movements and attention? To understand this issue, we have to be clear about what attention is. One of the more popular definitions of attention is Wil;iam James':

Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German. (James, 1890)

This definition is appealing and intuitive. At the same time, it is not in any obvious way related to eye movements. What can be taken from it, is that attention entails selection ("one out of what seem several simultaneously possible objects or trains of thought") and focusing of consciousness ("It is the taking possession by the

[4]The use of "saliency" vs. "salience" is - unfortunately - not consistent in this thesis. This is because the individual chapters were published before the rest of this thesis was written.

mind"). Later authors have focused on the selection aspect of attention and distinguish between covert and overt shifts of attention (Posner, 1980). Overt shifts of attention are orienting responses of the body towards a source of sensory input, while covert shifts of attention imply mental orienting (without movement) towards a source of sensory input. In James' definition covert attention cognitively selects one of several possible objects or trains of thoughts, while overt attention directs the sensory organs towards physical manifestations of these. In this scheme, eye movements qualify as overt shifts of attention.

The relation of overt and covert shifts of attention is intensely debated (Wright & Ward, 2008). Many researchers hold the view that eye movements are preceded or accompanied by a covert shift of attention (reviewed in Hoffman, 1998). Orientation discrimination performance is, for example, improved at saccade targets shortly before saccade onsets (Rolfs, Jonikaitis, Deubel, & Cavanagh, 2011). The saccade target appears to automatically attract attention and prohibits shifts of attention to other locations (Deubel & Schneider, 1996). Covert and overt shifts also use an overlapping set of neural structures (Nobre, Gitelman, Dias, & Mesulam, 2000; Corbetta et al., 1998). Furthermore, during covert shifts of attention, small eye movements (microsaccades) towards the direction of the attention shift can be observed (Laubrock, Engbert, & Kliegl, 2005; Hafed & Clark, 2002; Engbert & Kliegl, 2003). These findings suggest a tight relation between eye movements and shifts of attention.

In summary, most eye movements shift covert and overt attention to a different part of the environment. In this sense, eye movements are an integral part of attention. They select parts of the environment, just like covert attention does. Understanding where, when and how we move our eyes therefore inevitably informs us about attention.

1.3 What guides our eyes?

How can we explain and understand the process of selecting one location over another as a saccade target? On the one hand, we might ask what cognitive factors influence the selection of fixation locations and what the goals and purpose of fixation selection are. This allows us to formulate our answer relative to the goals, intentions and capabilities of the cognitive agent, i.e. in terms of psychological concepts. For example, we might explain that looking at the coffee mug facilitates grabbing it and allows us to achieve our goal of drinking the coffee. On the other hand, we might ask what the neural mechanisms for fixation selection are. In this case, our answer probably explains how retinal signals are relayed to the visual cortex, analyzed and prioritized until they eventually end up in saccade generating structures. Both approaches aim to provide an explanation for the same thing. Yet, both are distinct enough such that they can not easily be related to each other. For example, we do not know how goals, intentions, motivation and other psychological concepts can figure in neural mechanisms.

This is an important point because explanations of viewing behavior are often heavily biased towards one of such different styles of explanation. The bias is often caused by constraints imposed by recording techniques and the original research question[5]. On the one hand, neural explanations require the recording of neural signals (e.g. electroencephalography (EEG), functional magnetic resonance imaging (fMRI), single cell recordings). This more often than not implies that the subject's head must not move. This prohibits tasks that are set in natural environments. On the other hand, experiments in more natural environments often produce complex behavior which can not readily be described without reference to concepts such as intentions, goals and the task. On the upside, such weakly defined concepts allow the use of mathematical tools that can formalize and model complex behavior. An example of this is the application of

[5]Of course, the research question determines recording techniques and vice versa.

inverse reinforcement learning, where goals, values and rewards of the observer are inferred from behavioral data (Sprague & Ballard, 2003; Rothkopf, Ballard, & Hayhoe, 2007; Rothkopf & Ballard, 2013).

Accordingly, investigations of eye movement control can roughly be sorted along a continuum between cognitive and neural explanations. This implies that current answers to the question "What guides our eyes?" are also situated along the same continuum.

Investigations of eye movements during natural tasks often provide cognitive explanations. This is potentially because tasks constrain what information is needed to complete a task. Actions, for example, often require specific information (e.g. the location of the cup to be grasped, the speed of the ball to be hit) for successful completion. Consequently eye movement behavior is explained with reference to task constraints, and thereby with reference to cognitive concepts (see below for an overview of eye movement control under various tasks).

Removing complex and natural tasks from experimental instructions removes constraints that guide behavior. This disadvantage is compensated by tight control of the viewed stimulus in more artificial laboratory settings. During free viewing of pictures, for example, the visual stimulus is under tight control while exploration strategies of observers are not constrained. Complete knowledge of the visual input allows the comparison of eye movements to computational models of attention shifts (like the saliency map model), which predict overt shifts of attention. Computational models can therefore be used as tools that instantiate specific hypothesis about the control of eye movements. Many models of attention shifts instantiate hypotheses about visual processing in the brain to predict eye movements. Explanations based on such models therefore move towards neural explanations on the cognitive - neural continuum.

Consequently, computational models can be used to relate cognitive and neural explanations of viewing behavior by incorporating aspects of both explanations into one model. Many computational models can be seen as implementations of hypotheses derived from

functional analyses of cognitive systems. Such models provide a vehicle to formalize hypotheses, generate predictions, and allow us to study where predictions fail. Zednik and Jäkel (2014), for example, argue that Bayesian modeling generates hypotheses about how a system might solve a problem and how the required computations might be implemented - a conclusion that may well generalize to many other computational models. This resonates well with accounts of explanation that highlight the need for mechanistic multi-level explanations (Craver, 2007; Piccinini & Craver, 2011). Computational models are therefore a valuable tool for explaining the guidance of eye movements.

The presented articles in this thesis all revolve around a computational model of attention shifts: the saliency map model of attention (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998; Itti & Koch, 2000, 2001a). This thesis is thereby situated between cognitive and neural explanations. The manuscripts in this thesis provide tools to evaluate such models and investigate critical functional mechanisms and potential implementations in the brain. To this end, the work in this thesis contributes to the question "what guides our eyes", by providing detailed investigations of the saliency map model for shifts of attention.

The next section describes task driven eye-movement control which is (currently) out of the scope of the saliency map model of attention. This is done to give the reader a notion of what kind of behavior is not explained by this model. I then proceed to describe the saliency map model of attention in detail.

Task-driven eye movement control

One of the most important variables that controls eye movements is the observer's task. Already the first recordings of eye movements revealed that eye-traces during picture viewing are dependent on the task (Yarbus, 1967). Yarbus demonstrated that estimating the age of persons in an image lead to different traces compared to free viewing of the same image. The influence of tasks on eye movements

has consequently been studied in great detail.

Land and McLeod (2000), for example, show that cricket players employ sequences of eye movements that allow them to estimate the information needed to hit the ball. Players pursue the ball when it leaves the thrower's hand, then make a saccade to the anticipated bounce point of the ball and resume pursuit. Land and McLeod (2000) argue that the first pursuit allows prediction of the bounce point, while information at the bounce point allows estimation of the trajectory of the ball after the bounce point. Critically, since the ball bounces off an uneven surface, the trajectory of the ball before the bounce carries little information about the trajectory after the bounce. It is therefore not necessary for the cricket player to track the ball for the entire flight. Similar anticipatory saccades have also be found during table-tennis (Land & Furneaux, 1997) and squash (Hayhoe et al., 2012). On the other hand, baseball batters appear to pursue the ball as long as possible without anticipatory saccades to track the ball (Bahill & LaRitz, 1984), presumably because the entire trajectory allows inference about when to hit the ball. These findings suggest that eye movements strategies adapt to the task at hand to allow extraction of visual information necessary to complete the task.

Ballard and Hayhoe (1992) observed that eye movements preceded the location of the hand during a block-copying task. A task where observers where asked to construct a copy of a model system with individual blocks. During this task, the eyes often left locations before an action was finished and saccaded to locations that were important for the next action. Ballard and Hayhoe also found that observers frequently saccaded to the model that they were copying before they dropped a block, suggesting that making saccades is easier than memorizing the model. This is in line with results from change and attentional blindness experiments (c.f. Simons & Rensink, 2005; Simons & Chabris, 1999; O'Regan & Nöe, 2001). Interestingly, Flanagan and Johansson (2003) report that eye movement patterns between action execution and action observation are similar,

suggesting that both activate similar mechanisms.

These findings suggest that eye movements during sensory-motor coordination tasks are highly specific to the task and are a) tailored to carry out task-relevant actions and b) "cheaper" than memorizing the state of the world. This conclusion is also supported by studies including tasks like sandwich-making (Land et al., 1999), brick sorting (Triesch, Ballard, Hayhoe, & Sullivan, 2003), obstacle avoidance (Rothkopf et al., 2007) and driving (Johnson, Sullivan, Hayhoe, & Ballard, 2014). These findings highlight the tight integration between eye movements and actions.

A commonality between the tasks described above is that they have a sensorimotor structure that allows the definition of what information is needed to solve the task. A cricket player might need to estimate the speed, direction and general trajectory of the ball to hit it and making sandwiches consists of several actions that need to be carried out in a specific sequence. In the examples above, eye movements targeted objects and locations that allowed inferring certain attributes of the environment that were needed to solve the task (e.g. the speed of the ball). The fact that performance in a task is correlated with skillful execution of eye movement sequences suggests that task requirements strongly constrain potential eye movement targets.

However, there are many tasks where it is not known a priori to the observer which locations provide relevant information for solving the task. For example, in visual search paradigms observers search for a target stimulus embedded in the background. If the target is sufficiently similar to the background, the observer can not know with certainty if the next fixation location allows to detect the target. Since the target usually appears at each position with equal probability, the task does not constrain potential saccade targets by design. Yet, visual search strongly constrains eye movement sequences. Najemnik and Geisler (2005) carried out an ideal observer analysis of visual search for a faint Gabor patch embedded in pink-noise. Their ideal observer takes all available information into

account (like uncertainties introduced by the fall off of visual acuity) and carries out saccades that maximize the probability of finding the target. Comparing the ideal-observer to humans revealed that the minor differences between the two can likely be explained by sub-optimal information integration over the fixation history. The fact that an ideal-observer can, at least conceptually, be straightforwardly defined for visual search shows that visual search provides constraints on how to optimally plan eye movement trajectories and how to integrate information sampled during fixations into the planning process. Such visual search tasks do therefore not constrain trajectories by indicating what information can be sampled where, but the history of fixation locations and sampled evidence at those locations constrains possible future locations.

Other, more abstract tasks combine several kinds of constraints. Betz et al. (2010) found differences in viewing behavior of web pages depending on whether observers had to rate how relevant a web page was for a certain topic, or for which user group (e.g. 'gardeners') the site was relevant. Such tasks likely depend on the prior knowledge of the observer, of his/her prior exposure to websites and the ability to judge how relevant the page at hand would be for other users. Torralba, Oliva, Castelhano, and Henderson (2006) and Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) present a model that predicts saccade targets during search for people in real world scenes. The model incorporates context and task dependent spatial priors to predict saccade targets. In these tasks, both prior knowledge and the information sampled so far constrain future fixation locations.

In summary, the task carried out by an observer heavily influences eye movements by providing constraints for behavior. Specifically, skilled execution of a task often goes along with skillful and specific sequences of eye movements. Tasks constrain the information and the sequence of actions that are needed, and in the examples above eye movement patterns were seen to be tailored to provide exactly this information (c.f. Tatler, Hayhoe, Land, and Ballard, 2011).

Task-free eye movement control.

Given the task-specificity of eye movements, it is not surprising that some authors suggest that the task dominates the selection of fixation locations. This has led, for example, to formal models that attempt to deduce an observer's task strategy from eye movements (Rothkopf et al., 2007). Such studies point out a fundamental difficulty for task-based models of eye movements. It is presently unclear how the various tasks, which range from squash to making tea, can be formalized and conceptually integrated into models of the fixation selection process.

An alternative strategy to investigate eye movements is to focus on task-independent factors of eye movement guidance. One of the oldest "tasks" for eye-tracking studies is free-viewing of images (Yarbus, 1967). Observers are instructed to "freely" view pictures without an explicit task. Of course, it is unclear what it means to "just look" at an image. And with respect to the importance of the task, Tatler et al. (2011) have put forward the criticism that ignorance of the task does not imply the absence of a task. Free viewing is therefore best understood as a task that does not provide task-based constraints that are constant across observers. We (Wilming et al., n.d.) argue that free viewing samples from the prior distribution of possible tasks and that this is sometimes exactly what is desired. If task-based control is not at the center of interest, free viewing allows the sampling of many tasks to identify common factors between them. Free viewing therefore focuses on guiding factors that are present when the task does not constrain viewing behavior.

The existence of task independent factors, has been shown in several studies. Nuthmann and Henderson (2010) and Einhäuser, Spain, and Perona (2008), have shown that the presence of objects influences the selection of fixation locations. Firstly, objects are more often fixated than the background. Secondly, objects are usually fixated at very specific locations within the object borders. Kietzmann et al. (2011) have shown that during the viewing of ambiguous scenes gaze location biases perception towards one of the two possible

interpretations of the stimulus. Torralba et al. (2006) have demonstrated that the scene context (e.g. beach vs. indoors) influences eye-movements. Additionally, several occulomotor biases like the center bias (Tatler, 2007; Tatler & Vincent, 2009) and geometric structure in eye movement trajectories have been documented (Hooge, Over, van Wezel, & Frens, 2005; Tatler, 2007; Tatler & Vincent, 2009; Smith & Henderson, 2009, 2011b, 2011a; Wilming et al., 2013). These examples demonstrate that eye movements show distinctive patterns when the task provides no constraints for eye movements. The fact that free viewing elicits such consistent viewing behavior between observers (Wilming, Betz, Kietzmann, & König, 2011) suggests that free-viewing of pictures consistently activates similar mechanisms for the control of eye movements. Understanding the cognitive and neural underpinnings of these mechanisms is therefore important for understanding eye movement control in general.

1.4 The saliency map model of attention shifts.

How can we understand eye movements during free viewing of pictures? One of the most important concepts for explaining where we look is the saliency of stimulus locations. Saliency is an abstract quantity that expresses how much attention a location in the visual field attracts. In the context of covert attention, saliency is thought to express how likely a location will be attended next. In the context of overt attention it describes how likely a location will become the next saccade target. Saliency provides an abstraction of why a location should be looked at. For example, a stimulus might be salient because it stands out relative to its surrounding, or because it is highly relevant for the next action. Conceptually, saliency therefore provides a common currency that allows different processes to contribute to the distribution of attention.

With such an abstraction in place, predicting eye movements implies predicting the saliency of each location in a stimulus. The efficiency of such a saliency map can then be evaluated by comparing

it to the viewing behavior of observers. Crucially, different models of viewing behavior might produce different saliency estimates for a given location. Different models can therefore be compared by how well they predict fixation locations of human observers. Modeling viewing behavior by modeling saliency has become ubiquitous (Borji and Itti (2013) compare 64 (!) different models). Yet, the first formulation of a saliency map model of attention (Koch & Ullman, 1985) already contained, at least conceptually, most features of state of the art models that predict eye movements.

For this reason, I will start with Koch & Ullman's original work to introduce the saliency map model of attention with a historical perspective. Importantly, the Koch and Ullman (1985) model aimed at explaining covert shifts of attention. Itti et al. (1998), Itti and Koch (2000) later provided a computational implementation of the saliency map model that predicted covert and overt shifts of attention. The next sections therefore start with saliency for covert attention and then move on to saliency and overt attention.

Saliency and covert attention

Treisman and Gelade (1980) investigated reaction times during a covert attention search task. If the search target varied from the distractors in only one single feature (say color), search times where independent of the number of distractors. They suggested that in this ”pop-out” case search was carried out in parallel. When the target was defined by a conjunction of two features, search times were linearly related to the number of distractors. It thus seems as if observers had to inspect every item in the display in turn and check for the target. Treisman and Gelade (1980) suggested that conjunction search was carried out serially by directing a ”spotlight of attention” towards each item.

Koch and Ullman (1985) proposed that serial and parallel search can be explained by the same mechanism: the saliency map model of covert attention shifts. To this end they described a model for covert shifts of attention that rests on four different pillars. First,

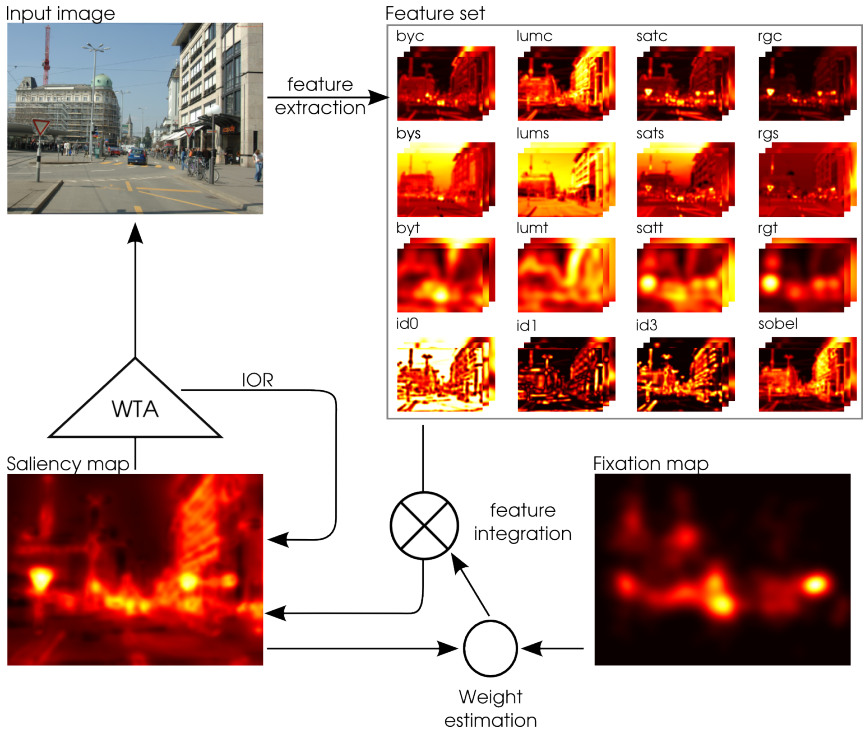


Figure 1.2: The typical structure of a saliency map model. Extracted features are from Wilming, Jutras, Buffalo, and König (n.d.). Feature integration combines single features into a saliency map. A winner-takes-all (WTA) mechanism with inhibition of return (IOR) determines the peaks of the saliency map. Each step has been implemented in many different ways by different models. As an example, we (Wilming, Jutras, Buffalo, & König, n.d.) use a logistic regression to determine optimal feature weights for linear feature integration.

they assume that the stimulus display is decomposed into simple and atomic feature maps. Each feature map encodes the value of a stimulus attribute for each stimulus location. Exactly what attributes are encoded by atomic features was not specified in detail, but features were suggested to be akin to the properties of visually receptive

cells in the early visual cortex. Different feature maps do not simply encode feature values for each location, but how "conspicuous", i.e. how distinct, each location is within each feature dimension. Importantly, these feature maps are retinotopically organized, which allows to compare locations across feature maps. Second, different conspicuity maps are combined into a single saliency map that globally encodes how conspicuous locations are across different conspicuity maps. Saliency maps therefore provide a retinotopic encoding of the overall conspicuity, or saliency, of locations. Third, a winner-take-all mechanism relays the content of the most salient location into a central representation (say working memory) for further processing. Fourth, the central representation consequently inhibits the most salient location to allow attention to shift and to ensure that previously visited locations are not immediately visited again (inhibition of return). Figure 1.2 shows a prototypical saliency map model that contains the basic elements of the Koch and Ullman (1985) model.

The original saliency map model of attention picks which location will be attended next, purely based on stimulus features. During Treisman & Gelades pop-out search the difference in saliency between distractors and target is large and the target is therefore picked immediately. During conjunction search, distractors and target are visually very similar and therefore similarly salient. The model needs to shift attention from item to item until the target is found. The saliency map model therefore successfully predicted the covert shifts of attention proposed by Treisman & Gelade.

To summarize, the original saliency map model focused on shifts of covert attention. The model shifts attention to the most salient location in a saliency map. Saliency is purely based on stimulus features and is computed before a location is attended. The original saliency map model therefore predicted covert shifts of attention and emphasized that saliency is determined by bottom-up stimulus features in a pre-attentive manner.

Covert attention models for overt attention

One of the surprising aspects of Koch and Ullman's saliency map model for covert attention shifts is that it became the standard model for overt shifts of attention as well. Itti and Koch (2000) proposed a computational implementation of the Koch and Ullman (1985) model that was able to generate predictions for attentions shifts on arbitrary images. This allowed to confirm that the model predicts search times for serial and parallel search tasks as found by Treisman and Gelade (1980). But more importantly, Itti & Koch also compared human behavior during search for military vehicles in large aerial pictures to performance of their computational model. Crucially, human observers were allowed to move their eyes during the search and the saliency map model was therefore compared to shifts of overt attention

Saliency map models of attention quickly became the standard approach for modeling overt shifts of attention. This can be seen by tracing how and when saliency map models were evaluated against eye movements to show their effectiveness. Figure 1.3 shows 21 saliency map models in their historical context. That is, after the publication of the Itti & Koch implementation of Koch & Ullman's model, a flurry of new models were published. These models improve the Itti & Koch model, but keep the overall structure (c.f. Borji & Itti, 2013). Strikingly, almost all models which appeared after 2000 evaluate how well they predict fixation locations (orange marker). This implies a conceptual move from modeling covert shifts of attention to modeling overt shifts of attention.

On the one hand this move seems plausible since we know that covert attention is related to overt attention (see section 1.2 above). On the other hand, overt shifts of attention pose serious problems for saliency models. During a saccade the visual field shifts such that the retinotopic map does not correspond to an egocentric map anymore. This implies that inhibition of return needs to take saccadic shifts into account. Matters are further complicated by the anisotropic distribution of rods and cones on the retina. The concep-

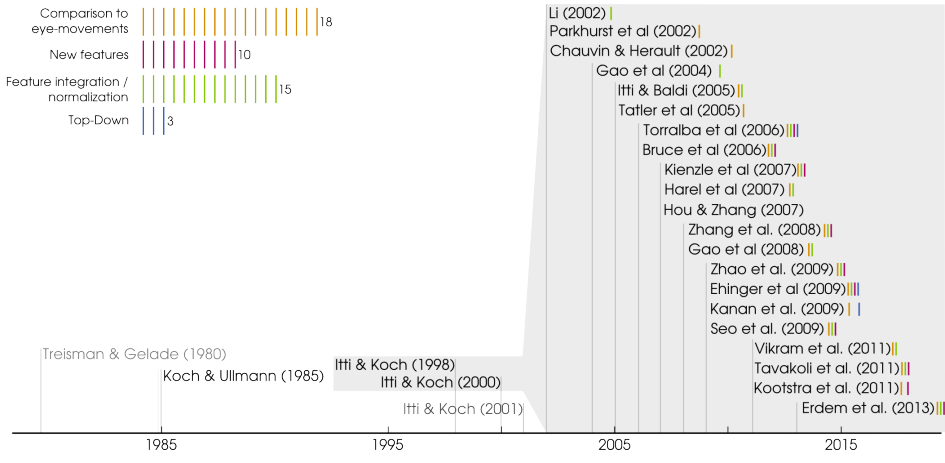


Figure 1.3: The development of saliency map models. The gray shaded area highlights 21 implementations of saliency map models. Each model is color coded to indicate advances relative to the Itti & Koch model. Orange: The model is evaluated against eye movement data. Purple: The model incorporates new visual features. Green: The integration of different conspicuity is different. Blue: The model incorporates top-down information. The top-left insets shows that almost all models after 2000 are evaluated against overt shifts of attention. Borji and Itti (2013) give a detailed summary of saliency models.

tual and algorithmic difficulties of these effects for the saliency map model have been recognized by Itti and Koch (2001a), but have not undergone detailed investigations.

Instead, authors focused on improving the prediction of saccadic endpoints with saliency maps. Several authors have compared Itti & Koch style saliency models to eye-tracking data (Chauvin & Herault, 2002; Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005). Some have centered their efforts on selecting features that should be included in the model (Frey, Honey, & König, 2008; Frey et al., 2011; Açik, Onat, Schumann, Einhäuser, & König, 2009; Jansen, Onat, & König, 2009; Onat, Açık, Schumann, & König, 2014; Kollmorgen, Nortmann, Schröder, & König, 2010; Einhäuser et al., 2008;

Kootstra, de Boer, & Schomaker, 2011). Yet others interpret locations that discriminate between objects as salient (Gao, Mahadevan, & Vasconcelos, 2008; Gao & Vasconcelos, 2004). Kienzle, Wichmann, and Scholkopf (2007), Zhao and Koch (2011) and Erdem and Erdem (2013) use machine learning techniques to combine features and weight image patches to predict fixation locations. Torralba et al. (2006) and Ehinger et al. (2009) integrate top-down information into the prediction. Another popular approach is to use Bayesian principles to integrate different features (Itti & Baldi, 2005a; Bruce & Tsotsos, 2006; Zhang, Tong, Marks, Shan, & Cottrell, 2008; Kanan, Tong, Zhang, & Cottrell, 2009; Seo & Milanfar, 2009; Tavakoli, Rahtu, & Heikkilä, 2011; Erdem & Erdem, 2013).

However, of the 64 saliency models reviewed by Borji, Sihite, and Itti (2013) none focuses on clarifying the relation between overt and covert models of attention (but Parkhurst et al., 2002 and Renninger, Coughlan, Verghese, and Malik, 2005 take the consequences of eye-movements into account). The bulk of the existing saliency map models predict saliency maps independent of the gaze location on an image. That is to say, most saliency map models ignore the distinction between overt and covert attention.

Yet, the improvements in predictive power have been dramatic. Judd, Durand, and Torralba (2012) provide an independent benchmark of a comprehensive list of saliency models[6]. In particular the best models achieve 76%[7] of the inter-observer consistency. The inter-observer accuracy expresses how well viewing behavior of individual observers matches that of a group of independent observers. Figure 1.4 shows the performance of a number of models evaluated by Judd et al as a function of publication year. This immediately shows how successful saliency map models are at predicting fixation locations of human observers.

[6] A more up to date list can be found online at http://people.csail.mit.edu/tjudd/SaliencyBenchmark/, last accessed July 2014

[7] I subtract 0.5 before taking the ratio, since an AUC value of 0.5 is chance performance

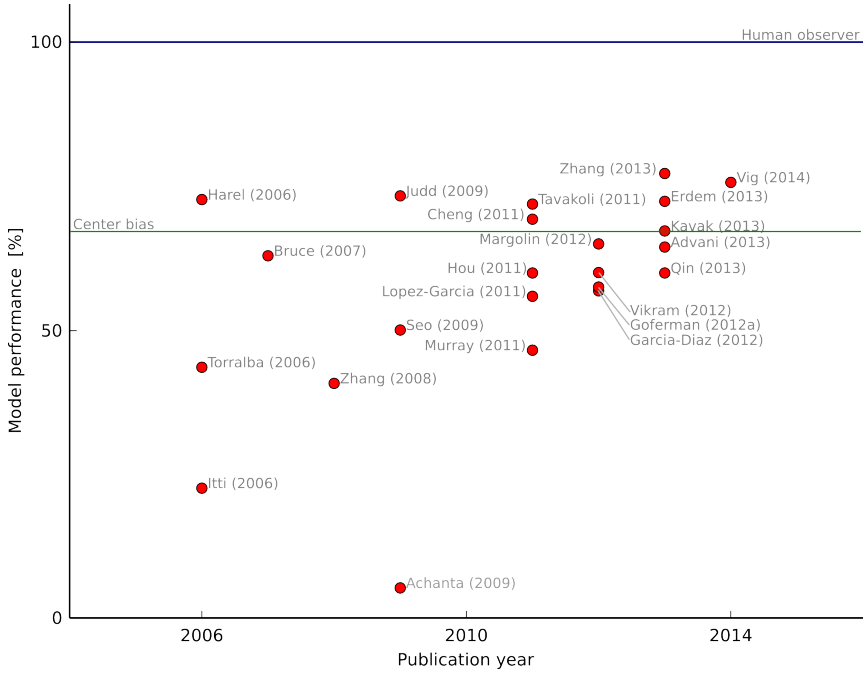


Figure 1.4: Development of saliency map model performance over time. Model performance is given relative to how well independent human observers predict other observers (blue line) and the predictive power of the center bias of fixations (green line). These form the reference frame that we suggest in chapter 2.

Importantly, these models are still closely related to the model proposed by Koch and Ullman (1985). In particular, all models decompose the image into some feature space, process these features and then integrate them into a single saliency map. They are therefore implementations of Koch & Ullman's saliency map hypothesis.

In summary, free-viewing of pictures produces consistent and systematic viewing behavior. This viewing behavior can, at least partly, be predicted by saliency map models of attention. This suggests that saliency and the saliency map model of attention are important tools

for the understanding of mechanisms that guide eye movements.

Yet, how much we can learn from a computational model about the cognitive and neural underpinnings of the guidance of eye movements is not clear a priori. On the one hand, some models predict fixation locations of human observers very well. This argues for the importance of the saliency map model. But if predictive power is an arbiter for the importance of saliency map models, then it is necessary to ensure that measures of predictive power are valid. On the other hand, saliency map models ignore the retinal consequences of overt shifts of attention. Whether or not static saliency maps can generate realistic eye movements by saccading to the most salient location and inhibiting previous locations is an open question. The saccade generating components of the saliency map model must therefore be investigated in detail. Chapters 2 and 3 pick up these topics and investigate different measures of predictive power and the inhibition of return component of saliency map models.

1.5 Attention in the brain

A model that is a good explanation of viewing behavior must predict fixation locations well and veridically explain how the brain selects saccade targets. Whether or not the saliency map model does this is an intensely debated question. The model itself is not specific enough to only allow one biologically plausible implementation. Consequently, many different implementations are possible and debated. The brain might use, for example, a retinotopically organized brain area, whose neurons encode saliency, to select saccade targets. Alternatively, eye movements might be controlled by a network of areas where each area competes for selecting the next saccade target. To untangle different possible implementations, the saliency map model must be put side by side to what is known about mechanisms of fixation selection in the human brain. The goal of the next paragraphs is to provide a brief summary of different brain areas that have been implicated in fixation selection. The focus of this

summary is on how the processing in these areas might relate to the saliency map model of attention.

The neural basis of the saliency model

Koch and Ullman (1985) suggested that the saliency map is either located early in the visual hierarchy (lateral geniculate nucleus (LGN) or primary visual cortex (V1)) or later beyond areas like medial temporal area (MT)/medial superior temporal area (MST) and visual area 4 (V4). Eventually, several additional areas have been suggested as the locus of the saliency map, namely V4, lateral intraparietal sulcus (LIP), frontal eye field (FEF), superior colliculus (SC) and the pulvinar. Here I briefly review and list studies that support the role of these areas in the control of overt and covert attention (in many cases during visual search comparable to Treisman and Gelade, 1980). In particular I concentrate on six properties that are essential for saliency maps (w.r.t. the definition by Koch & Ullman, 1985; Itti & Koch, 2001a). (i) A saliency map integrates over many visual features and its activation is therefore feature independent, i.e. saliency is a common currency that allows different mechanisms to guide attention. (ii) The peak of a saliency map encodes the object to be attended next. (iii) Inhibition of already visited locations leads to lower activity for already explored locations. (iv) A saliency map is retinotopically organized. (v) It is independent of motor control, i.e. a saliency map can lead to covert and overt shifts of attention. (vi) A saliency map integrates top-down influences. Top-down influences are defined in contradistinction to the influence of stimulus features ("bottom-up"). Typical examples for top-down influences are goals, prior-knowledge, expectations and context. The next paragraphs relate findings in particular brain areas to these criteria.

Frontal eye field Several studies have shown that some FEF neurons respond to visual stimulation independent of visual features (i): FEF neurons discriminate between targets and distractors during easy visual search within ~130ms (Thompson, Hanes, Bichot, & Schall,

1996), reliably with few neurons (Bichot, Thompson, Chenchal Rao, & Schall, 2001), independent of the stimulus type (color vs. motion) (Sato, Murthy, Thompson, & Schall, 2001), and the time-course of target-distractor discrimination predicts response times (Bichot et al., 2001). Causal evidence for FEF's role during search is provided by Wardak, Ibos, Duhamel, and Olivier (2006), who show that chemical inactivation of FEF leads to prolonged response times during search. Further evidence that implicates FEF in visual search comes from Monosov and Thompson (2009). They report that trial-by-trial activity of target selective FEF neurons is positively correlated to behavioral reports during visual search (ii), i.e. strong FEF activity can erroneously lead to the selection of a distractor. Also, Thompson, Bichot, and Sato (2005) demonstrate that saccades to distractors can be traced back to FEF activity reaching levels usually found before saccades to targets. Both findings suggest that the most active neuron in FEF selects the next saccade target. With respect to inhibition of return (iii), Bichot and Schall (2002) demonstrate that a target shown at the same location in consecutive trials is discriminated slower by FEF neurons. This is expected if previously attended locations are inhibited. FEF receives topographically organized afferents (iv) from areas in the ventral and dorsal stream (MT, TEO, V4, visual area 3 (V3), visual area 2 (V2), LIP; Schall & Morel, 1995). In analogy to the saliency map model, FEF therefore might integrate different visual feature information from lower level areas. Importantly, FEF activity is also present during covert attention tasks (v). Thompson, Biscoe, and Sato (2005) show that saccade related neurons are inhibited when eye movements are irrelevant for the search process. And Schall and Morel (1995) find target-distractor discriminating activity even in NOGO trials where no eye movements are required. Given that stimulation of FEF can elicit saccades (Schlag-Rey, Schlag, & Dassonville, 1992; Fujii, Mushiake, & Tanji, 1998), FEF can potentially support covert and overt attention. FEF activation also appears to be susceptible to top-down influences and FEF can additionally itself exert top-down influence over other areas. That FEF can exert

top-down influence over other areas was shown by three studies. Moore and Armstrong (2003) demonstrate that sub-threshold stimulation of FEF neurons enhances activity in spatially corresponding units in V4. Monosov, Sheinberg, and Thompson (2011) argue that inactivation of FEF indirectly leads to decreased activity of inferior temporal cortex (IT) neurons when their preferred object is in their receptive field (RF). Premereur, Vanduffel, and Janssen (2014) find that FEF can modulate LIP firing rates under specific conditions. Top-down modulation of FEF activity has also been demonstrated (vi). Bichot, Schall, and Thompson (1996) show that extensive training leads to feature selectivity of FEF units at the expense of 'popout' detection. Thompson, Bichot, and Sato (2005) argue that errors after feature switching in popout search tasks are based on top-down priming of FEF for the wrong target. The interpretation that FEF can exert top-down influence and receives top-down information was recently corroborated by findings from Buschman and Miller (2007) and Bastos, Vezoli, and Bosman (2014). Buschman and Miller (2007) investigated the relation of area LIP and FEF. They found that LIP responds quicker to bottom-up saliency (i.e. pop-out targets) than FEF, and FEF reacts quicker to top-down saliency (i.e. search targets) than LIP. The effects of bottom-up saliency and top-down saliency were correlated with changes in oscillatory coupling between FEF and LIP. Bottom-up saliency lead to stronger synchrony in the slow Gamma range, and top-down saliency increased synchrony in the fast beta range. Recently Bastos et al. (2014) found that feedforward projections are mediated by gamma-band synchronization and feedback projections by beta-band synchronization. Taken together the two studies suggest that LIP relays bottom-up information to FEF and FEF sends top-down information to LIP. The importance of oscillatory activity for shifts of attention in FEF was furthermore strengthened by Buschman and Miller (2009), who found that serial shifts of attention were correlated with 18-34Hz oscillations in the local field potential of FEF.

In summary, there is a large body of evidence that connects FEF to

the processing of saliency during visual search. Neurons in FEF can quickly discriminate between targets and distractors independent of feature dimensions. FEF is furthermore retinotopically organized and stimulation of FEF neurons elicits eye movements. At the same time, FEF receives top-down information and can send out top-down information to other areas. It is therefore not surprising that many authors have suggested that FEF implements a saliency map (see Thompson and Bichot, 2005 for a review).

LIP Several lines of investigation have provided evidence that the lateral intraparietal area encodes saliency in a retinotopic fashion. LIP is often called a "priority map" instead of a saliency map to highlight the importance of top-down signals (see for example Bisley and Goldberg, 2003; Arcizet, Mirpour, and Bisley, 2011a). For simplicity reasons I will use the term saliency map with the understanding that it incorporates top-down signals. LIP's retinotopic organization (iv) has been demonstrated by several studies (Blatt, 1990; Swisher, Halko, Merabet, McMains, & Somers, 2007; Jerde, Merriam, Riggall, Hedges, & Curtis, 2012). (Gottlieb, Kusunoki, & Goldberg, 1998) have shown that LIP neurons react to visual search targets when they are behaviorally relevant or become salient due to a sudden onset. Arcizet et al. (2011a) show that neurons in LIP respond to the saliency of stimuli, independent of their feature dimension (i). The activity of LIP neurons is closely related to saccades (Ipata, Gee, Goldberg, & Bisley, 2006; Thomas & Paré, 2007; Ipata, Gee, Bisley, & Goldberg, 2009; Premereur, Vanduffel, & Janssen, 2011), but LIP neurons do not always trigger saccades (Ipata et al., 2009; Arcizet et al., 2011a) (v). Mirpour, Arcizet, Ong, and Bisley (2009) and Mirpour and Bisley (2012) argue that activity of LIP neurons is greatest for potential future targets (ii), less for already visited targets (iii) and even less for distractors during visual search. Neurons in LIP furthermore integrate top-down information (vi, because they react in a task specific manner (Gottlieb et al., 1998; Jerde et al., 2012)) and differentiate targets and distractors (Ipata, Gee, Gottlieb, Bisley, &

Goldberg, 2006; Thomas & Paré, 2007; Ipata et al., 2009; Mirpour et al., 2009). Inactivation of LIP interferes with performance in visual search tasks (Wardak, Olivier, & Duhamel, 2004).

These findings show that LIP carries all the properties that would be expected of a priority map. The conclusion that LIP indeed acts as a saliency (or priority) map has been put forward by many authors, for reviews see Bisley and Goldberg (2010), Bisley (2011).

Superior Colliculus The superior colliculus is a structure in the midbrain (Krauzlis, Lovejoy, & Zénon, 2013) that is important for the control of eye, reach and head movements. In other animals it is important for the control of other orienting responses such as sonar vocalization, pinnae and whisker movements (Gandhi & Katnani, 2011). The superior colliculus is retinotopically organized (iv) (Krauzlis et al., 2013) and microstimulation of the superior colliculus can trigger saccades towards the receptive field of the stimulated neuron. Kustov and Robinson (1996) demonstrated that such elicited saccades are shifted towards the locus of covert attention, which suggests that the superior colliculus is more than a simple motor control structure. Furthermore, activity of some superior colliculus neurons is related to the probability that a saccade is made into the RF (Basso & Wurtz, 1998; Dorris & Munoz, 1998; Horwitz & Newsome, 2001). These studies clearly establish the superior colliculus as an important part of attentional eye movement control. The superior colliculus shows additional properties that would be expected of a saliency map. Like neurons in LIP and FEF, neurons in the SC show selective and discriminatory activity for targets in their receptive fields (i) (Glimcher & Sparks, 1992; Krauzlis & Dill, 2002; McPeek & Keller, 2002). But superior colliculus activity is also present for perceptual tasks where no eye movements are required (v) (Horwitz, Batista, & Newsome, 2004) or at times well before a saccade (McPeek & Keller, 2002). Sapir, Soroker, Berger, and Henik (1999) report of a patient with unilateral damage to the superior colliculus who shows less inhibition of return in the affected visual field (iii).

Visuomotor neurons in the superior colliculus also show smaller target/distractor discrimination in a classical Posner task when the target is cued vs. uncued (Dorris, Klein, Everling, & Munoz, 2002; Fecteau & Munoz, 2005). Furthermore (ii), microstimulation of the SC leads to perceptual improvements akin to covert attention (Cavanaugh, Alvarez, & Wurtz, 2006; Cavanaugh & Wurtz, 2004; Müller, Philiastides, & Newsome, 2005) and influences selection of saccade targets (Carello & Krauzlis, 2004; Dorris, Olivier, & Munoz, 2007). Conversely chemical inactivation of parts of the SC leads to impairments of covert and overt shifts of attention (Lovejoy & Krauzlis, 2010).

All of the cortical areas that have been suggested to be a saliency map project to the SC (vi): FEF (Künzle, Akert, & Wurtz, 1976), LIP (Gaymard, Lynch, Ploner, Condy, & Rivaud-Pechoux, 2003), V4 (Gattass, Galkin, Desimone, & Ungerleider, 2014) and V1 (Wurtz & Albano, 1980). And Bell and Munoz (2008) argue that activity in the SC reflects attentional effects due to top-down and bottom-up selection strategies. The SC therefore either combines top-down and bottom-up information or receives the combination of these signals from an upstream area. It is also known that the SC relays an efference copy of movement signals to cortical areas via the pulvinar (Gilbert & Li, 2013). But Zénon and Krauzlis (2012) show that inactivation of the SC does not diminish attention effects in cortical areas MT/MST. The SC therefore does likely not send top-down information to cortical areas.

In summary, the superior colliculus is strongly involved in covert and overt shifts of attention. Its central location within the oculomotor system and extensive connectivity with cortical areas make it a prime candidate for the integration of diverse attention signals, and thereby a prime candidate for a saliency map.

Pulvinar The role of the pulvinar for covert and overt attention shifts has been less intensely studied compared to LIP, FEF or the superior colliculus. Lesions or chemical inactivation of the pulvinar

lead to deficits in visual attention (Ungerleider & Christensen, 1979; Robinson & Petersen, 1992). Furthermore, LaBerge and Buchsbaum (1990) identified the pulvinar, in a positron emission tomography (PET) study, as contributing to an object identification task when the task demanded spatial attention. Some pulvinar neurons selectively react to stimuli that signal the target of upcoming saccades or when covert attention is directed toward them (v) (Petersen, Robinson, & Keys, 1985). Fischer and Whitney (2012) show that the pulvinar only encodes attended objects and filters out distractors. The pulvinar contains several retinotopic maps (iv) of the visual field (Petersen et al., 1985; Robinson & Petersen, 1992) that receive extensive inputs from areas V1, V2, V4, MT, TEO, TE, FEF and LIP (i, reviewed in Robinson & Petersen, 1992; Shipp, 2004; Grieve, Acuña, & Cudeiro, 2000). These results allow for the possibility that the pulvinar might compute saliency by combining information from many cortical areas.

Furthermore, Purushothaman, Marion, Li, and Casagrande (2012a) have shown that the pulvinar can strongly suppress or boost V1 activity in a retinotopic fashion, effectively gating which information is processed by V1. Saalmann, Pinsk, Wang, Li, and Kastner (2012) show that the pulvinar synchronizes alpha activity according to attentional demands in V4 and TEO.

Taken together, these results suggest that the role of the pulvinar is to combine cortical signals to compute saliency. Its output is subsequently used to select which stimuli are processed by the cortex.

V4 Some studies have implicated area V4 in the processing and computation of saliency. V4 is a retinotopically organized (iv) area involved in, amongst others, color, shape and form processing (Roe et al., 2012) (i). Mazer and Gallant (2003) show that V4 neurons have high activity before a saccade is made into their RF during visual search (ii). They furthermore report that activity of V4 neurons is related to visual processing, but not to saccade control (v). Mazer

and Gallant (2003) also find that V4 neurons show target selectivity, probably due to top-down effects (vi). These findings are corroborated by studies from Bichot, Rossi, and Desimone (2005), Ipata, Gee, and Goldberg (2012) and Melloni, van Leeuwen, Alink, and Müller (2012). Whether inhibition of return plays a significant role in V4 has to my knowledge not been investigated (iii).

These studies clearly implicate area V4 in the deployment of attention and suggest that it has a role for the computation of saliency. Given that V4 is a lower level visual area (Felleman & Van Essen, 1991; Markov et al., 2014), it remains to be seen whether V4 is best described as a saliency map or as a feature computation unit.

V1 Koch and Ullman (1985) have suggested that the saliency map might reside in early visual areas. Li (2002) argues that V1 provides a saliency map to downstream areas. Zhaoping and May (2007) find that a model of V1 (Li, 1998, 1999b) replicates effects of pop-out search and conjunction search (i,ii). Further evidence for saliency processing in V1 comes from Zhang, Zhaoping, Zhou, and Fang (2012). They report that pop-out stimuli that are too faint to be consciously perceived still produce behavioral attention effects. The magnitude of these effects is correlated with recorded EEG and blood oxygen level dependent (BOLD) signals that were localized in V1. V1 also shows other properties expected from a saliency map. It has a retinotopic organization (iv) and activity is independent of saccade control (v). Whether V1 receives attentional top-down feedback is discussed. On the one hand Melloni et al. (2012) find that V1 is not influenced by top-down information (vi). On the other hand Brown and Guenther (2012) speculate that IOR might act by feedback from ventral stream areas to V1 that leads to stronger IOR for objects (iii). It is also known that V1 has reciprocal connections with many other visual areas (for example V2, V3 and V4, Felleman and Van Essen (1991)) and that V1's activity is directly shaped by projections from the pulvinar (Purushothaman, Marion, Li, & Casagrande, 2012b) and indirectly by the cerebellum (Sultan et al., 2012).

In summary, primary visual cortex possesses many of the attributes expected of a saliency map. It should be pointed out though, that the "standard" view of V1 processing is more that of a non-linear filter stage (Carandini et al., 2005). Whether such non-linear filters are sufficient to compute saliency on natural stimuli is unknown.

A network phenomenon?

The above paragraphs show that many areas in the brain exhibit characteristic properties of a saliency map. Clearly, the fact that the saliency map model contains only one saliency map is at odds with the abundance of saliency processing areas.

However, the saliency map model proposes a functional hierarchy of areas that can potentially be mapped to the cortical hierarchy. In this view, different feature / conspicuity maps might be implemented by different areas. This view and the abundance of saliency signals in different brain areas has lead many authors to conclude that the computation of saliency is not confined to one area, but is probably expressed in a network of areas (Serences et al., 2005; Fecteau & Munoz, 2005; Ipata et al., 2009; Bisley, 2011; Jerde et al., 2012; Melloni et al., 2012; Anton-Erxleben & Carrasco, 2013; Krauzlis et al., 2013; Shipp, 2004).

We thus have to ask: Can the network of saliency processing areas be mapped to the saliency map model of attention?

A definite answer to this question is not known. However, what is known about different saliency processing areas allows to make some informed suggestions. As a first step, anatomical projections between areas can be used to constrain possible mappings between saliency processing areas and the saliency map model of attention. Such projections are directed and thereby allow to order different areas into a processing hierarchy that can be compared to processing steps in the saliency map model of attention.

Recently Markov et al. (2014) mapped the cortical hierarchy by measuring feed-forward and feed-back connections between many different visual areas. Their hierarchy places V1 at the bottom, V4

and parts of FEF (responsible for small saccades) at an intermediate level, followed by LIP and finally the rest of FEF (large saccades). Anatomically defined hierarchies therefore constrain the roles of different areas as feature, conspicuity or saliency maps.

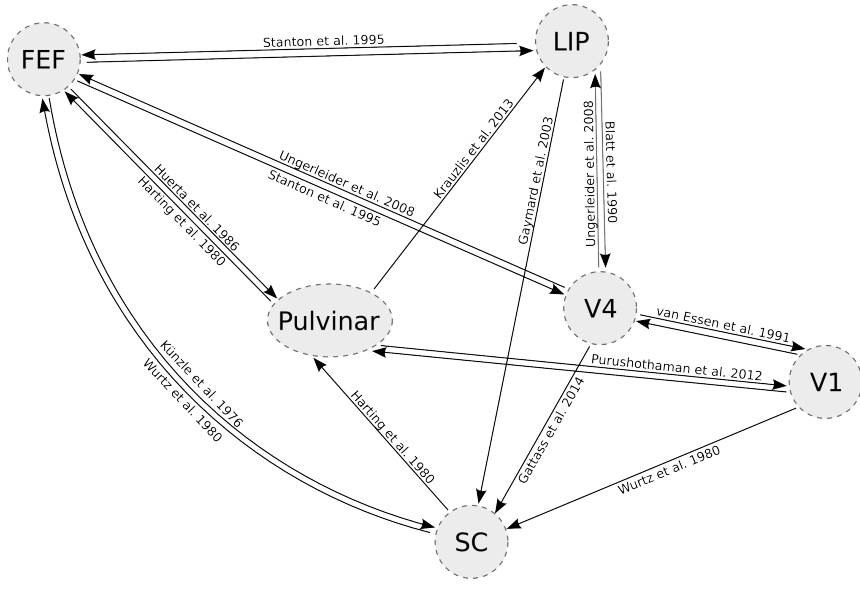


Figure 1.5: Brain areas implicated in the processing of saliency. Arrows between areas indicate direct connections, labels indicate publications that show existence of connection.

The anatomical hierarchy fits to some extent to the functional roles of areas proposed by the studies reviewed above. Figure 1.5 summarizes the connections between areas that have been implicated in the processing of saliency. In particular, several researchers have suggested that FEF and LIP, which are the highest areas, are saliency maps that can directly control attention. In this scheme LIP might integrate conspicuity signals from lower level areas, like V1 and V4, and projects saliency signals to FEF. The difference between FEF and LIP would then be one of different abstractions, with FEF potentially receiving top-down information from pre-frontal cortex.

Alternatively, Shipp (2004) suggests that the pulvinar is the locus of the saliency map. In this view, the pulvinar pools information from visual cortex (e.g. V1-V2,V4, TEO & TE), computes saliency and biases processing in visual cortex towards salient stimuli. FEF and LIP provide top-down information and inhibition of return via the superior colliculus.

However, while such suggestions seem plausible, we are currently unable to rule out alternative mappings of the saliency map of attention to the brain (Shipp, 2004 alone compares seven different proposals). A major obstacle in relating the saliency map model to brain areas is the strong connectivity between different areas. When one area provides feed-back to another, it is difficult to distinguish between areas that simply show attentional modulation and areas that encode saliency. An area that receives top-down feedback from a saliency map will likely process salient stimuli differently from non-salient stimuli and thus potentially appear as if it computes saliency itself. Top-down modulation from a saliency map might therefore be mistaken as saliency computation.

This problem becomes clearer when we look at how the normalization model of attention (Reynolds & Heeger, 2009) incorporates attentional signals. The normalization model of attention expresses how neurons within one area are affected by attentional signals and by lateral connections. At the same time, the normalization model of attention is able to explain diverse effects of attention onto single unit activity (contrast gain: Reynolds, Pasternak, & Desimone, 2000; response gain: Williford & Maunsell, 2006; competition between stimuli: Martínez-Trujillo & Treue, 2002; Treue & Trujillo, 1999; Reynolds & Desimone, 2003; multiplicative scaling of tuning curves: McAdams & Maunsell, 1999; sharpening of tuning curves by feature attention: Martinez-Trujillo & Treue, 2004). The model assumes that covert attention selectively increases the gain of neurons responding to specific locations or feature dimensions. Output firing rates are then determined by dividing the firing rate by the suppressive drive that it receives from neighboring neurons (w.r.t. receptive field loca-

tion and feature dimension). Boosting some locations and feature dimension (i.e. because of top-down feedback) automatically leads to suppression of other locations and feature dimensions. In this model, top-down feedback can therefore shape large scale activity in brain areas by boosting salient locations.

Given that we accept the normalization model of attention as a fair description of the consequences of attention, we must therefore also accept the possibility that top-down feedback from a saliency map can turn any retinotopically organized brain area into an area that looks like a saliency map (because salient locations are boosted) even though it does not compute saliency.

Importantly this possibility is not only restricted to feed-back information from a saliency map. In particular, it might also be the case that an early area computes saliency and higher levels simply receive saliency modulated feed-forward information from this area.

In summary, to identify a unique saliency map it is necessary to untangle the individual contributions of several areas. As detailed above, this is a difficult undertaking that needs to differentiate between areas that compute saliency and areas that receive saliency modulated input. On the one hand, if saliency is computed early upstream areas might appear like saliency maps but do not actually compute saliency. On the other hand, lower level areas might receive saliency modulated top-down feedback from higher areas (Buffalo, Fries, Landman, Liang, & Desimone, 2010). In this light the suggestion that V1 acts as a saliency map is especially important because almost all other areas receive feed-forward input from V1. Chapter 4 will return to this topic by investigating the contributions of V1 to saliency computation.

1.6 About this thesis

The work contained in this thesis investigates how eye movement targets are selected. The individual manuscripts in this thesis address specific questions about different parts of the saliency map

model of attention.

Study one: How should attention models be evaluated?

In the previous sections I have argued that it is difficult to map the saliency map model of attention to the brain. Furthermore, the model is commonly used to predict overt shifts of attention even though it was originally devised for covert shifts of attention. This raises the question of whether or not the saliency map model is a good model for explaining viewing behavior.

In the context of what we know about the visual system and eye movement selection, it became apparent that knowledge of visual processing mechanisms only weakly constrain models of saliency computation. This is evident in the large number of models that are usually related to the visual system and the lack of concrete biological models. Given insufficient anatomical and functional constraints for what models are good ones, the predictive power of models is an important proxy for what models are good explanations. At the very least, we require that good explanations should generate models that predict empirical data well.

It is however, not immediately obvious what it means to be a good predictor of human eye movement data. The evaluation process requires the choice of appropriate measures, the consideration of oculomotor biases and an evaluation of how much variance to be predicted there is in a dataset. Chapter 2 and the accompanying publication (Wilming, Betz, Kietzmann, & König, 2011) investigate this topic in great detail.

This part of the thesis therefore deals with evaluating the effectiveness of saliency map models. It aims at enabling better comparisons between models to aid in separating good from bad models.

Study two: Is inhibition of return an essential component of free viewing behavior?

An essential component of saliency map models of attention is inhibition of return. However, whether inhibition of return saccades exists is questionable. The original phenomenon of inhibition of return has been established with covert shifts of attention. Whether oculomotor inhibition of return exists is far from clear. Some groups argue for the existence of oculomotor inhibition of return (Klein & MacInnes, 1999) and others against it (Hooge et al., 2005; Smith & Henderson, 2009, 2011a, 2011b).

Clearly, if inhibition of return does not exist in the oculomotor system, the saliency map model must be adapted and inhibition of return must be replaced with a more plausible saccade generation mechanism.

In chapter 3 we[8] show in a large study with over $\frac{1}{2}$ million fixations that inhibition of return does not adequately describe the fixation selection process. Our data therefore does not support the assumption that inhibition of return exists in the form assumed by saliency map models. On the contrary, we find that those locations that are fixated more than once, are more salient in a bottom-up sense. That is those locations that should be inhibited are returned to more often than locations that are presumably not inhibited.

Study three: Does V1 implement a saliency map?

Koch and Ullman (1985) hypothesized that very early areas like the LGN or V1 implement a saliency map. This seems to be in contradiction with suggestions about the locus of the saliency map from numerous other studies. But since V1 is directly connected to many other areas, and to even more indirectly, V1 would also relay the saliency signal to these areas. It is therefore possible that saliency signals in other higher areas are merely a consequence of

[8]We refers to my coworkers and myself.

saliency processing in V1. Whether or not V1 contains a saliency map has therefore consequences for the interpretation of the role of many other areas that contain saliency signals.

Chapter 4 presents an fMRI experiment in which we investigate if early visual areas encode saliency by dissociating the processing of saliency from the processing of luminance contrast. We found that BOLD activity in early visual cortex is largely dependent on luminance contrast, but not on saliency. Multivariate decoding techniques could not decode saliency signals from early visual brain areas. These findings speak against the hypothesis that early visual areas compute saliency.

Study four: Guidance by saliency in monkeys and humans

Most of what we know about the neural basis of oculomotor control and the function of early visual cortex is based on studies in animals, most prominently with macaque monkeys. Specifically, much of the knowledge that led to the formulation of the biological underpinnings of the Koch and Ullman (1985) saliency model comes from studies in macaque monkeys. This begs the question if oculomotor control structures are the same in humans and monkeys, and if viewing behavior is comparable between both species. Surprisingly, only few studies have directly compared viewing behavior of humans and monkeys on stimuli that are also used for evaluating saliency models.

Chapter 5 provides such a comparison. We compared viewing behavior of four macaque monkeys with that of 108 human observers on the same set of stimuli. Our results indicate that bottom-up saliency influences are virtually identical in both species. However, humans and monkeys often fixate different locations, even on stimuli that are ecologically valid for neither species. This suggests that top-down influences are different between species. The comparable influence of saliency in both species suggests that macaque monkeys are a good model system for bottom-up influences, but likely less useful for other models of eye movement guidance.

Study five: Beyond saliency.

I hope that the preceding sections have made clear that the role of the saliency map model in an active vision setting is not clear yet. Specifically, some authors doubt that the saliency map model of attention can be adapted to explain eye movements during active everyday behavior (Tatler et al., 2011). One of the fundamental problems here is that we have not fully understood the what early visual areas do (Carandini et al., 2005). That is to say, it is unclear what information (e.g. objects, sensory consequences of actions?) could potentially contribute to the computation of saliency. The last chapter (6) takes a step back and sketches rough outlines of a theory that could lead to the development of alternative accounts of overt shifts of attention.

In summary, the five different manuscripts investigate different levels of the saliency map model of attention. In the second chapter, I investigate how we can evaluate whether or not a model predicts eye movement behavior well. The third chapter focuses on inhibition of return, which is a critical component of the saliency map model of attention. A part of the neurological basis of the saliency map model is investigated in the fourth chapter. The fifth chapter compares viewing behavior between humans and monkeys to assess whether saliency plays a similar role for both species. Last, I present some thoughts and ideas that might lead to the development of models that go beyond feature driven saliency.

Chapter 2

Measures and Limits of Models of Fixation Selection

This chapter has been published in "PLoS ONE":
Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and Limits of Models of Fixation Selection. *PLoS ONE*, *6*(9), e24038

2.1 Abstract

Models of fixation selection are a central tool in the quest to understand how the human mind selects relevant information. Using this tool in the evaluation of competing claims often requires comparing different models' relative performance in predicting eye movements. However, studies use a wide variety of performance measures with markedly different properties, which makes a comparison difficult. We make three main contributions to this line of research: First we argue for a set of desirable properties, review commonly used measures, and conclude that no single measure unites all desirable properties. However the area under the ROC curve (a classification measure) and the KL-divergence (a distance measure of probability distributions) combine many desirable properties and allow a meaningful comparison of critical model performance. We give an analytical proof of the linearity of the ROC measure with respect to averaging over subjects and demonstrate an appropriate correction of entropy based measures like KL-divergence for small sample sizes in the context of eye-tracking data. Second, we provide a lower bound and an upper bound of these measures, based on image-independent properties of fixation data and between subject consistency respectively. Based on these bounds it is possible to give a reference frame to judge the predictive power of a model of fixation selection. We provide open-source python code to compute the reference frame. Third, we show that the upper, between subject consistency bound holds only for models that predict averages of subject populations. Departing from this we show that incorporating subject-specific viewing behavior can generate predictions which surpass that upper bound. Taken together, these findings lay out the required information that allow a well-founded judgment of the quality of any model of fixation selection and should therefore be reported when a new model is introduced.

2.2 Introduction

A magnificent skill of our brain is its ability to automatically direct our senses towards relevant parts of our environment. In humans, the visual capacity has by a large margin the highest bandwidth, making directing our eyes towards salient events the most important method of selecting information. We sample the visual input by making targeted movements (saccades) to specific locations in the visual field, resting our gaze on these locations for a few hundred milliseconds (fixations). Controlling the sequence of saccades and fixation locations thereby determines what parts of our visual environment reach our visual cortex, and contingently conscious awareness. Understanding this process of information selection via eye movements is a key part of understanding our mental life.

A common approach to investigate this process has been to use computational models that predict eye movements to gain insights on how the brain solves the problem of determining where in a scene to fixate (Itti & Koch, 2001b; Itti & Baldi, 2005b; Kanan et al., 2009; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Parkhurst et al., 2002; Peters, Iyer, Itti, & Koch, 2005; Zhang et al., 2008). The similarity of empirical eye-tracking data and model predictions is then used as an indication of how well the model captures essential properties of the fixation selection process. For this chain of reasoning, i.e. for drawing inferences about the workings of the brain, it is highly relevant how the quality of a model of fixation selection is measured. Furthermore, if different models are to be compared and judged, there needs to be an agreed upon metric to make this comparison possible. Of equal importance for model comparisons is the data set that is being used as ground truth. Different data sets might be more or less difficult to predict, which confounds a potential model comparison across different studies. In this article, we investigate metrics for evaluating models of fixation selection, and methods to quantify how well models of fixation selection can score on a specific data set. This leads to a framework for evaluating and comparing

models.

Before we can discuss how measures and data set influence the evaluation, we have to be clear about what models of fixation selection actually predict. Even though the ultimate goal of the model may be to predict fixation locations, the actual mechanism of fixation selection is usually not addressed in detail. Instead the focus is on computing a topographic representation of how strongly different parts of the image will attract fixations. Classically, each region in an image is assigned a so-called salience value based on low-level image properties (e.g. luminance, contrast, color) (Itti & Koch, 2001b; Itti & Baldi, 2005b; Kanan et al., 2009; Kienzle et al., 2009; Parkhurst et al., 2002; Peters et al., 2005; Zhang et al., 2008). The topographic representation of the salience values for all image regions is known as the salience map. Some models furthermore incorporate image-independent components, like the fact that observers tend to make more fixations in the center of a screen than in the periphery regardless of the presented image, known as a spatial (or central) bias (Tatler & Vincent, 2009; Zhang et al., 2008; Tatler, 2007; Tatler et al., 2005). Other forms of higher level information that have been used in models of fixation selection are task-dependent viewing strategies, information about face-locations and search-target similiarity (Cerf, Harel, Einauser, & Koch, 2008, 2009; Hwang, Higgins, & Pomplun, 2009; Torralba et al., 2006). However, even in those models the important output is a map of fixation probabilities. Thus, in accordance with the focus on this approach in the modeling literature, we restrict our analysis to the evaluation of models that generate a salience map. Since the empirical data that these salience maps have to be evaluated against are not maps themselves, but come in the form of discrete observations of fixation locations, it is not obvious a priori how to judge the quality of such a model.

In the first part of this article, we therefore review different commonly used evaluation measures. We define properties that are desirable for evaluation measures and provide evidence that many commonly used measures lack at least some of these properties. Because

no single measure has all of the desirable properties, we argue that reporting both the Area Under the receiver-operating-characteristic Curve (AUC) for discriminating fixated from non-fixated locations, and the Kullback-Leibler divergence (KL divergence) between predicted fixation probability densities and measured fixation probability densities, gives the most complete picture of a model's capabilities and facilitates comparison of different models.

In the second part of this work, we turn to properties of fixation distributions and examine what impact they have on model evaluation and comparison. Our aim is to formalize the notion of how difficult a data set is to predict, which will facilitate comparisons between models that are evaluated on different datasets. We use the image- and subject-independent distribution of fixation locations (spatial bias) to establish a lower bound for the performance of attention models that predict fixation locations. The predictive power of every useful model should surpass this bound, because it quantifies how large evaluation scores can become without knowledge of the image or subject to be predicted. Complementary to this, we use the consistency of selected fixation locations across different subjects (inter-subject consistency) as an upper bound for model performance, following Ehinger et al. (2009), Cerf et al. (2009), Einhäuser et al. (2008), Harel, Koch, and Perona (2007), Hwang et al. (2009), Kanan et al. (2009). The reliability of these bounds depends on how well they can be estimated from the data being modeled. We therefore provide a detailed investigation of the spatial bias as well as inter-subject consistency, and their dependence on the size of the available data set. This establishes a reference frame that allows judging whether improvements in model performance are informative of the underlying mechanism and facilitates model comparison.

Finally, we examine the conditions under which the proposed upper bound holds by turning to a top-down factor that has so far been neglected in the literature. We show that incorporating subject idiosyncrasies improves the prediction quality over the upper bound set by inter-subject consistency. This should be interpreted as a

note of caution when using our proposed bounds, but does not call into question their validity in the more general and typical case of modeling the viewing behavior of a heterogeneous group of subjects.

2.3 Results

Measures of model performance

In this section, we review commonly used measures for the evaluation of models of fixation selection. Our aim is to investigate, on a theoretical basis, what the advantages and disadvantages of different measures are and to identify the most appropriate measure for model evaluation. To reach this aim, we choose a four step approach. First, we establish a list of desirable properties for evaluation measures. Second, we identify commonly used measures in the literature and describe how they compare model predictions to eye-movement data. Third, we assess how the measures fare with regard to the desirable properties. Justified by this, we recommend the use of the AUC. Finally, we elucidate the effect of pooling over subjects and conclude that in some circumstances, KL-divergence is a more appropriate measure.

Desirable Properties for evaluation measures Evaluation scores of a model of fixation selection will at some point be used to compare it to other models. Such comparisons are not only difficult because different data sets are being used, but also because the interpretation of evaluation measures can be difficult. Informed by our own modeling work and by teaching experience, where several points repeatedly obstructed the comparison of different models, we define two properties that help to interpret evaluation scores:

- Few parameters: The value of an evaluation measure ideally does not depend on arbitrary parameters, as this can make the comparison of models difficult. If parameters are needed, meaningful default values or a way of determining the parameters

are desirable.

- Intuitive scale: A good measure should have a scale that allows intuitive judgment of the quality of the prediction. Specifically, a deviation from optimal performance should be recognizable without reference to an external gold standard.

Models of fixation selection are usually evaluated against eye-tracking data, which is typically very sparse in relation to the size of the image that is being viewed. It is therefore desirable for an evaluation measure to give robust estimates based on low amounts of data:

- Low data demand: During a typical experiment, subjects can usually make only a relatively small number of saccades on a stimulus. Thus, an ideal measure should allow for a reliable estimate of the quality of a prediction from very few data points.
- Robustness: A measure should not be dominated by single extreme or unlikely values. Consider, for example, that the prediction of a fixation probability distribution consists of potentially several million data points. The result of the prediction of a single data point should not have a large impact on the overall evaluation. A measure should also be able to deal with the kinds of distributions typically occurring in eye-tracking data. A fixation density map (see Materials and Methods: Fixation density map estimation) is usually not normally distributed but, due to its sparseness, dominated by the presence of many very unlikely events.

The properties presented here aim at ensuring that an evaluation measure is suitable to deal with eye-tracking data and to ensure that an evaluation score can be meaningfully interpreted. The list is not necessarily exhaustive, but we argue that any exhaustive list would have to contain these properties.

Existing measures To identify commonly used measures, we sought articles that present or compare salience models which operate on

static images of natural scenes. We used the Google Scholar bibliographic database (scholar.google.com) to search for articles that were published after the year 2000 and contain the words "eye", "movement", "model", "salience", "comparison", "fixation", "predicting" and "natural" somewhere in the text. This list of key-words was selected because omitting any one of them disproportionately increases the number of results unrelated to models of human eye movements. The search was performed on June 28, 2011. We manually checked the first 200 articles for evaluations of salience models on static natural scenes. In the resulting 25 articles (Hwang et al., 2009; Itti & Baldi, 2005b; Kienzle et al., 2009; Peters et al., 2005; Itti & Koch, 2001b; Torralba et al., 2006; Açik et al., 2009; Cerf et al., 2009; Harel et al., 2007; Baddeley & Tatler, 2006; Elazary & Itti, 2008; Betz et al., 2010; Butko & Movellan, 2008; Ehinger et al., 2009; Parkhurst et al., 2002; Einhäuser et al., 2008; Cerf et al., 2008; Zhang et al., 2008; Kanan et al., 2009; Renninger, Verghese, & Coughlan, 2007; Bruce & Tsotsos, 2009; Kootstra et al., 2011; Yanulevskaya, Marsman, Cornelissen, & Geusebroek, 2011; Parikh, Itti, & Weiland, 2010; Tatler et al., 2005; Tatler & Vincent, 2009) eight different measures are used to compare eye-tracking data to predictions of fixation locations.

We sort the seven different measures into three groups, based on the comparison they perform. The three measures in the first group, chance-adjusted salience, normalized scan-path salience and the ratio of medians, compare the central tendency of predicted salience values at fixated locations with salience values at non-fixated locations. The second group, comprising 80th percentile, AUC and the naïve Bayes classifier, treats the salience map as the basis for a binary classification of locations as either fixated or non-fixated and evaluates the classification performance. The third group includes the KL-divergence and the Pearson product moment correlation coefficient. For these measures, the model output is interpreted as a fixation probability density, and the difference between this and a density estimated from actual fixation data is computed.

- *Chance-adjusted salience (S_a)* (Parkhurst et al., 2002) is the difference between the mean salience value of fixated locations on an image and the mean salience value of the viewed image. Thereby, if values are larger than 0, salience values at fixated locations are above average.
- *Normalized scan-path salience (NSS)* (Peters et al., 2005) is the mean of the salience values at fixation locations on a salience map with zero mean and unit standard deviation.
- The *ratio of medians* (Parikh et al., 2010) compares the salience values at fixated locations to the salience at random control points. The salience value of a location is determined by finding the maximum of the salience map in a circular area of radius 5.6 degree around that location. The median salience at fixated locations and the median salience of a set of random control points on the same image are computed for each image. The ratio of both medians is used as evaluation measure.
- The 80^{th} *percentile measure* (Torralba et al., 2006) reports the fraction of fixations that fall into the image area that is covered by the top 20% of salience values. It therefore reports the true positive rate of a classifier that uses the 80^{th} percentile of the salience distribution as a threshold. The selected area covers, by definition, 20% of the image, which is therefore the expected value for a random prediction.
- The *area under the receiver-operating-characteristics curve (AUC)* (Tatler et al., 2005) describes the quality of a classification process. Here, the classification is based on the salience values at fixated and non-fixated image locations. All locations with a salience value above a threshold are classified as fixated. The AUC is the area under the curve that plots the true positive rate against the false alarm rate for all possible thresholds (the receiver operating characteristic). As the threshold is continuously lowered from infinity the number of hits and false alarms are both increasing. When the salience map is useful, the hits will increase faster than the false alarms. With still lowering

threshold the latter will catch up and the fraction of hits and false alarms both reach 1 (100%). The AUC gives an estimate of this trade-off. An area of 1 indicates perfect classification, 100% hits with no false alarms. An area of 0.5 is chance performance. See Fawcett (2006) for an introduction to ROC analysis.

- The *percent correct of a naïve Bayes classifier* (Tatler & Vincent, 2009) that distinguishes between salience values at fixated and non-fixated locations can be used as a model evaluation measure. The classifier is trained by estimating the probability distributions $P(S|F)$ and $P(S|\overline{F})$, where S refers to the salience value of a point and F signals if the point was fixated or not, on a subset of the data. Unseen data points are classified as fixated based on their salience if $P(F|S) > P(\overline{F}|S)$. The percent correct score is computed in a cross-validation scheme such that all data points are classified as part of the test set once.
- The *Kullback-Leibler divergence (D_{KL})* (Itti & Baldi, 2005a; Itti & Baldi, 2005b) is a measure of the difference between two probability distributions. In the discrete case it is given by:

$$D_{KL}(P||Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)})$$

In the case of salience map evaluations, P denotes the true fixation probability distribution and Q refers to the model's salience map that is a 2D probability density function. For every image location the true fixation probability is divided by the model fixation probability and the logarithm of this ratio is weighted with the true fixation probability of the location. Therefore, locations that have a high fixation probability are emphasized in the D_{KL} values. The D_{KL} is a non-symmetric measure ($D_{KL}(P||Q) = D_{KL}(Q||P)$ does not hold for all P and Q). This is irrelevant for model evaluation, but becomes relevant when it is not clear what the true probability is, e.g. for evaluating inter-subject variability. In this case, a symmetric extension of D_{KL} can be obtained by $D_{KL}(P||Q) + D_{KL}(Q||P)$.

- The *Pearson product-moment correlation coefficient (correlation)* (Kootstra et al., 2011; Hwang et al., 2009) is a measure of the linear dependence between two variables. The correlation coefficient between two samples is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  where X and Y are the two variables, and $\bar{X}$ and $\bar{Y}$ are the sample means. Evaluating models of fixation prediction with this measure requires a little conceptional gymnastics. If the values in a prediction map are interpreted as observations of variable X, and the values in the empirical fixation probability distribution at the same pixel locations are interpreted as observations of variable Y with the same index, the correlation coefficient between prediction and ground truth can easily be computed. The Pearson product moment correlation coefficient is bounded between -1 for predictions that are the inverse of the ground truth (ground truth multiplied with a negative number, plus or minus any number), and 1 for perfect predictions. A value of 0 indicates that there is no linear relation between the prediction and the empirical fixation density.

Evaluation of measures with respect to the described properties Having proposed a list of desirable properties and introduced a number of different measures, we can now examine how these measures cope with the requirements and what aspect of the prediction they evaluate. For an overview, please see Table 2.1.

- *Few parameters:* There are three measures that do not have parameters: S_a, NSS and AUC. The ratio of medians is dependent on the radius that is used for selecting a salience value for a fixation. Although there may be reasons for choosing one value over another, this parameter is essentially arbitrary. The percentile chosen for the 80^{th} percentile measure is completely

arbitrary; it might as well be the 82^{nd} percentile. For the naïve Bayes classifier , the correlation and the KL-divergence, it is necessary to estimate probability distributions, which in the simplest case depends on the binning used. The naïve Bayes classifier furthermore requires the specification of the number of cross-validation runs.

- *Intuitive scale:* S_a does not have an intuitive scale since the mean and range of a salience map are arbitrary and both influence the scale. The ratio of medians method is also not intuitive as it is not obvious how the resulting scores are to be interpreted. What does it mean that salience at fixated locations is 1.3 times higher than at random locations? What would it mean if it were 1.4 times higher instead? The interpretation of KL-divergence scores is also difficult for similar reasons. NSS has a rather intuitive scale because it uses the standard deviation of the salience map as its unit. All three classifying measures (80^{th} percentile, AUC, naïve-Bayes) are bounded, which should make their score easy to interpret by comparing the model score to the theoretical maximum. However, when using eye-tracking data, the categorization of points into the classes 'fixated' and 'non-fixated' is non-trivial. Strictly speaking, there are no non-fixated points: If we just record data long enough, there is no principle reason why a specific point on the screen cannot be fixated. Thus, any method for selecting non-fixated and fixated points will produce overlapping sets, which cannot be perfectly separated. In turn, no classifier can reach its theoretical maximum score in this task. In Materials and Methods:Theoretical maximum value for AUC we show how to approximate the actual theoretical maximum score of the AUC, given a set of fixations. Despite these considerations, the meaning of classification performance (80^{th} percentile, naïve Bayes) is straightforward. The meaning of the AUC is not as intuitive but also allows to quickly assess the quality of a model. The interpretation of correlation scores is rather intuitive: scores are bounded from both sides and can

be interpreted as the linear dependence between prediction and ground truth. However, interpretation of a specific correlation value becomes less trivial if the actual dependence structure is not linear. In that case, which is typical for fixation data, the measure can be misleading when interpreted as if the condition of linearity was fulfilled.

- *Low data demand:* The three methods that require probability density functions, KL-divergence , correlation and naïve Bayes classifier, require a lot of data to form accurate estimates of the necessary probability distributions. In contrast, all other methods use only the fixated locations as positive instances and can in principle be computed on very few data points.
- *Robustness:* S_a uses the mean to summarize information about salience values at fixation locations. Since the mean is not robust against outliers, neither is S_a. NSS also uses the mean, but first normalizes the salience map to zero mean and unit standard deviation. Thus, extreme outliers will have a weaker effect than for S_a, but still influence the result. The ratio of medians uses the median as a descriptive statistic of salience at fixated and control points. This ensures that extreme outliers have no negative effect. The naïve Bayes classifier is not by definition robust against outliers, as its robustness depends very much on how the necessary probability distributions are estimated. If simple bin counting is used it is not robust against outliers. Similar arguments hold for the KL-divergence and the correlation, where the true fixation probability distribution has to be estimated from the data.

In summary, our evaluation shows that there are large differences in the suitability of the different measures when it comes to evaluating models of fixation selection. S_a, NSS and the ratio of medians are not intuitive to interpret and/or not robust. From the three classification measures, the AUC appears to be most favorable. It improves on the 80^{th} percentile measure by removing the arbitrary parameter and by including false alarms into the analysis. The naïve

Bayes approach needs more data than is often available and the estimation of probability density maps is non-trivial. Correlation and KL-divergence need much data and require the estimation of density functions. Additionally, KL-divergence is not easy to interpret, but has a sound theoretical basis when the comparison of probability densities is concerned. The AUC stands out on the properties we have outlined. Based on our defined requirements, the AUC seems to be the best choice for evaluating models of fixation selection.

The effect of pooling over subjects The selection of an appropriate measure is only one aspect of the evaluation process. Additionally, properties of the data against which the model is evaluated are of importance. Usually, when devising models of fixation selection, we are interested in the combined viewing behavior of several subjects, i.e. fixation data is pooled across subjects. The model should preferably predict those locations that are fixated by many subjects, because these fixations are most likely caused by salience or other factors that are stable across subjects, and not causes of fixations that are irrelevant to understanding information selection mechanisms. As a consequence of this, models that are trained to predict the joint-subject viewing behavior should perform better in predicting fixations from a set of subjects than in predicting the individual subjects from that set. This important property of model quality is not captured by the AUC and NSS.

Figure 2.1 shows an example, where the quality of prediction as measured by AUC or NNS for the combined smooth fixation density map is just as good as the average quality of prediction of the individual subjects. That this is a general property of the NSS measure is easy to see: it takes the mean salience values at fixated locations, and for the mean it does not make a difference whether we take it for subsets individually and then average over the resulting value, or take the mean of the complete set directly. The linearity of AUC under decomposition of positive observations into subsets is less obvious, but proven in Materials and Methods: Proof of AUC linearity. In

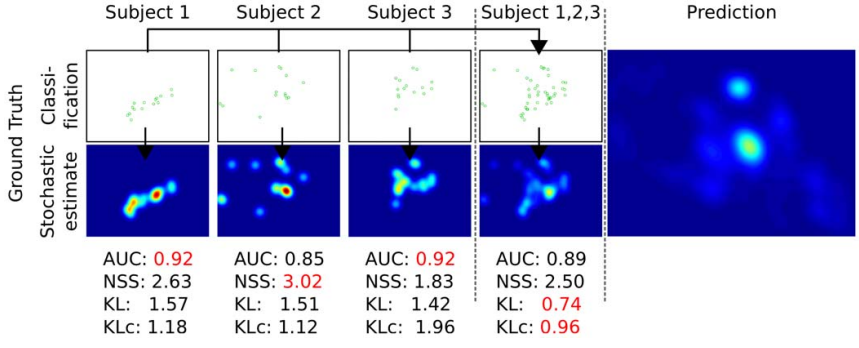


Figure 2.1: **P**redicting the joint fixation selection process of several subjects vs. predicting individual subjects. The prediction in this case was generated not from a model but from the fixations of several independent subjects. It therefore captures the joint process of a group of subjects. When treated as a classification problem (top row), only the fixation locations are important. In this case, the mean of the AUC or NSS scores for the individual evaluations are identical to the AUC or NSS score of evaluating the joint process. When treated as a stochastic process (bottom row; see Materials and Methods: Fixation density map estimation for computational details of fixation density map estimation), locations that were fixated by one but not all subjects are less important to predict. KL-divergence, which evaluates not individual fixations but the prediction of the stochastic process, yields a better score for the evaluation of the joint process. This also holds true when it is corrected for the number of fixations in the data (KLc).

contrast, KL-divergence and correlation yield better values for predicting the joint viewing behavior, because they operate on fixation density map estimates, which take the spatial relation between fixations into account, and they are thus able to give non-linearly more weight to those locations that have been looked at by many subjects (see Figure 2.1). This non-linear weighting can be a good reason to consider the KL-divergence or correlation for model evaluation, despite their computational difficulties mentioned above. Deciding which of the two measures to use when one wants to exploit the effect of pooling over subjects is a difficult questions. Both measures are not robust, and both have the potentially disadvantageous property of being sensitive to non-linear monotonic transformations of the prediction. Correlation has the advantages of boundedness and being slightly less sensitive to some rescalings of the model output. However, the intuitive interpretation of its scale breaks down and becomes misleading if the dependence that is being measured is not really linear. KL-divergence is extremely sensitive to low (close to zero) predictions for locations that get a higher empirical salience, but is conceptually more appropriate for comparing probability distributions. In the end, both measures are not optimal, but because of its sound theoretical basis, we recommend using the KL-divergence when one wants to capture the ability of the model to exploit similarities in the viewing behavior within a group of subjects. In practical applications of this measure, one should also be aware of an additional complication: KL-divergences are dependent on the number of fixations used to compute the fixation density maps (Materials and Methods: Fixation density map estimation). As a result, values which are estimated from different numbers of fixations are not directly comparable. For example, when the average fixation duration in an experiment with fixed viewing time per stimulus is dependent on image category, this can confound a comparison between categories. In Materials and Methods: Correction of KL divergence for small samples we investigate this dependency and describe a method for correcting KL-divergence scores for the bias introduced

by limited data by exploiting the measure's relation to information entropy. In summary, the linearity of AUC under decomposition into subsets and the sensitivity of KL-divergence and correlation for joint-viewing versus single-subject behavior are both relevant whenever a model of fixation selection is evaluated against fixation data. KL-divergence is especially appropriate when fixation data from a group of subjects are the target of a prediction.

Intermediate Summary This section focused on a theoretical investigation of different evaluation measures that are used to evaluate models of fixation selection. We conclude that AUC excels with respect to our list of desired properties: The disadvantage of non-intuitive interpretation of the meaning of the AUC is outweighed by it's non-parametric nature, boundedness, robustness and compatibility with small sample sizes. In practice, it is often useful to average evaluation scores across subjects and images in order to reduce the variance introduced by small sample sizes. The linearity of the AUC ensures that these averages retain a meaningful interpretation. This property, however, comes at a cost. When the goal is to predict consistent fixation behavior across all subjects, more weight should be given to locations that are consistent between observers. Here we recommend the use of the KL-divergence. However, it is important to employ algorithms that minimize a systematic bias in the case of few data points available (see Materials and Methods: Correction of KL divergence for small samples).

Properties of fixation data

The aim of the second part of this work is to investigate the upper and lower bounds on the prediction performance of fixation selection models. To this end, we examine the image and subject independent spatial bias on the one hand, and image-specific inter-subject consistency on the other hand. We use data from an eye tracking study carried out previously in our group (see Materials and Methods: Description of the eye-tracking study for details and Figure 2.2 for some

Figure 2.2: **F**our representative exemplary stimuli from each category used in the eye-tracking study. The top row shows natural scenes, the bottom row shows examples from the urban scenes. The right-most panels depict the spatial distribution of the first 15 fixations across all 64 images and 48 subjects in the two categories. On the natural scenes, there is a rather strong central fixation bias, while on the urban images fixations are more spread out.

examples of stimuli). We first analyze what kind of predictions can be achieved purely from the spatial bias without any knowledge of the image that is being viewed, and evaluate how this lower bound is influenced by the number of subjects and images available for its estimation. Secondly, we describe a method for computing an upper bound for model performance that is based on 'inter-subject consistency' and investigate in how far it depends on the number of subjects used for its computation.

The upper and lower bounds are based on predictions blind to the predicted subject. Notably, the inter-subject consistency ignores subject idiosyncrasies. The question thus arises whether the upper bound proposed here is really an absolute upper bound for the predictive power of models of fixation selection. We therefore investigate firstly whether knowledge of the subject idiosyncrasies can be utilized to improve predictions, and secondly whether we can combine image- and subject-specific information to surpass the upper bound given by the inter-subject consistency.

Estimating the lower bound for fixation selection models A way to estimate the lower bound for performance of fixation selection models is to compute the predictive power of the spatial bias. This prediction does not exploit information specific to the image or subject whose fixations are being predicted. Thus it has to be surpassed by any valuable model of fixation selection. Here, we take into account that the spatial bias varies between different image classes (Figure 2.2). We estimate the lower bound for NSS and AUC as the best representatives of central tendency measures and classification measures. As the results for AUC and NSS are qualitatively very similar, only the former is further considered here. More details on NSS results can be found as reference values in Materials and Methods: Reference values for spatial bias and inter-subject consistency. Since we explicitly wish to consider small data sets, KL divergence is not suitable here (but see Materials and Methods: Reference values for spatial bias and inter-subject consistency). To obtain a better understanding of the reliability of the lower bound, we investigate the dependence of the estimation quality on the number of subjects and images used. Specifically, we compute the lower bound by predicting fixation patterns of one subject on one image (the test set) with fixation data from other subjects on other images (the training set). To predict fixations in the test set, we construct an FDM from the training set and interpret it as a prediction for fixations in the test set. To quantify the quality of this prediction, we compute the AUC and NSS between the calculated FDM and fixations in the test set. To assess the dependence of the spatial bias estimation quality on data set size, we vary the number of images and subjects used to create the FDM. In detail, we individually increase the number of subjects and images in the training set exponentially from 1 to the maximum in seven steps ($N_{img} \in \{1, 2, 4, 8, 16, 32, 63\}$; $N_{sub} \in \{1, 2, 4, 7, 13, 25, 47\}$). For each of the 49 combinations, we use every image and subject combination as the test set 47 times such that each of the repetitions is one random sample of images and subjects for the training set. To avoid using specific subject-image combinations more often than others,

we treat cases in which we draw only one or two images or subjects separately. In this case the training set is explicitly balanced over repetitions and different test sets. In the other cases the large number of possible combinations ensures a roughly even sampling. We report the predictive power of the spatial bias as the mean over test subjects, test images and repetition.

The spatial bias depends on the image category (Figure 2.3, naturals and urban scenes left and right respectively, $p < 0.0001$). Furthermore, an increasing number of subjects (Figure 2.3, rows of large matrix, $p < 0.0001$) and images (Figure 2.3, columns of large matrix, $p < 0.0001$) significantly increase the predictive power of the spatial bias estimate (three factorial ANOVA, category X number of subjects X number of images). For natural scenes (left) the increase is steeper than for urban scenes (right) and thereby suggests that eye-movement patterns across subjects and stimuli are more similar during the viewing of natural scenes. The predictive power of the spatial bias estimate reached for the maximum number of subjects is surprisingly high (AUC of 0.729, 0.673 for naturals and urban scenes respectively) and poses a challenging lower bound for prediction performance. The predictive power of the spatial bias estimate increases extremely slowly when more than 32 images and 25 subjects are used, implying that the estimation becomes reliable at this point. A smaller number of subjects can be compensated by a larger image set and vice versa. However, using too few data leads to a danger of underestimating the lower bound and thereby overestimating one's model quality. In conclusion, the reliability of the lower bound estimation depends on the size of the data set; for all practical purposes, 32 images and 25 subjects seem to be sufficient for a reliable estimate.

Estimating the upper bound for fixation selection models To derive the upper bound for fixation selection models, we estimate the inter-subject consistency analogously to the spatial bias reliability. The rationale is that, due to variance across subjects, models that do

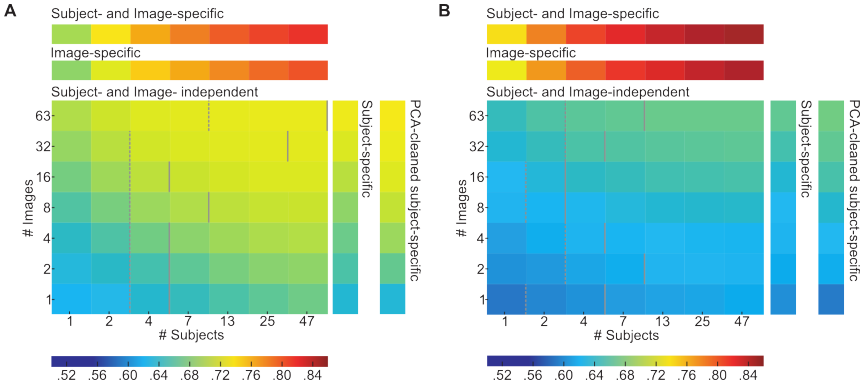


Figure 2.3: **E**stimation of lower and upper bounds for natural (A) and urban scenes (B). All data shown are AUC values averaged over all predictions of single subjects on single images in a given parameter combination. The predictions are based on a spatial bias (large matrix, 'Subject and Image independent'), a subject-specific bias (column next to the matrix, 'Subject-specific'), a PCA-cleaned subject-specific bias (rightmost column), an image-specific bias (row above the matrix, 'Image-specific', also referred to as inter-subject consistency) and the combination of image and subject-specific bias (topmost row). The 'Subject and Image independent' scores depend on the number of subjects and images used for the prediction and represent a lower bound for fixation selection models. The 'Image-specific' scores also depend on the number of images and yield an upper bound for fixation selection models. Comparing 'Subject-specific' and Subject and Image independent reveals the effect of using a subject-specific bias. The dashed lines indicate at what subject group size the subject-specific bias stops being significantly better than the spatial bias (paired t-test, $p > 0.05$). The subject-specific bias is not significantly different from the spatial bias between the dashed and solid lines. See main text for more detailed descriptions.

not account for individual idiosyncrasies cannot perform perfectly. Therefore, comparing model scores to a score obtained by predicting fixations from one subject with other subjects provides an intuitive normalization. If the model score and inter-subject consistency are equal, the model predicts a new subject's fixations as well as other subjects' fixations would. In the following, we investigate the dependence of inter-subject consistency on the number of subjects used for the prediction. To estimate inter-subject consistency, we first separate subjects into a test and a training set and compute an FDM from the training set. Then, we measure how well this FDM predicts the one subject in the test set. In contrast to above, the images in test and training sets are identical. To obtain a maximally accurate estimate of the training set size for which inter-subject consistency saturates, the number of subjects in the training set is increased in steps of one. Similar to the procedure above, we use every subject and image combination 47 times as test set for every possible number of subjects in the training set. For each of the 47 repetitions a random set of training subjects is drawn. The cases where only one or two training subjects are drawn are explicitly balanced across test subjects. In the following, we report the mean AUC over test subjects, test images and repetitions as a measure of inter-subject consistency. As expected, the inter-subject consistency increases with the number of subjects in the training set (Figure 2.3, second row from top 'image-specific' in panels A and B, $p < 0.0001$; one factorial ANOVA with number of subjects as factor; additional datapoints omitted for clarity). With the maximum number of training subjects, AUC is 0.802 for naturals and 0.846 for urbans. In contrast to the pure spatial bias predictions, predictability is higher for urbans than for naturals. This results in a dynamic range of the AUC between lower and upper bound of 0.073 and 0.173 for naturals and urbans respectively. Looking at the development of inter-subject consistency with increasing subject set size, it is reasonable to assume that further increasing the training set would not have a strong effect. The second derivative of the curve is always negative, suggesting that the curve

saturates. For example, from 20 to 21 subjects, the increase is 0.001, from 40 to 41 it is only 0.0002. Thus, for all practical purposes, the inter-subject consistency of about 20 subjects constitutes an upper bound for generic models of fixation selection in free viewing tasks.

Subject-specific spatial bias To investigate the importance of subject idiosyncrasies for the prediction of fixation locations, we examine whether knowledge of a subject-specific spatial bias is more valuable than knowledge of the bias of other subjects. To that end, we estimate how well a subject-specific spatial bias predicts fixations of the same subject on other images. We proceed as before and predict fixations in the test set with an FDM based on fixations in the training set. For every combination of the number of predicting images, test subject, and test image, we use 63 different training sets. The images in the different training sets are randomly sampled and the subject is the same in training and test set. The random samples are balanced explicitly if there are only one or two images in the training set. Analogous to the generic spatial bias, the subject-specific spatial bias's predictive power is dependent on the number of images used for estimation (Figure 2.3; vertical bar 'subject-specific' directly to the right of the large matrix in panel A and B, $p < 0.001$, ANOVA with number of images as the only factor). For any number of images, the subject-specific spatial bias is more predictive than the predictive power of a single independent subject (Figure 2.3 compare left-most column in the central square to vertical column directly to the right). However, it is not higher than the predictive power of the best spatial bias, obtained from a set of 47 independent subjects (Figure 2.3 compare right-most column in the central square to vertical column directly to the right). With the exception of 63 images from the 'natural' category, the bias from a large number of subjects achieves better performance than the subject-specific bias. The exact number of subjects that is needed to achieve better performance than the subject-specific spatial bias depends on the number of images (see dashed lines in Figure 2.3). The improvement in AUC over a generic

prediction based on a single independent subject ranges from 0.009 on urbans and 0.021 on naturals for a single image to 0.017 on urbans and 0.029 on naturals for 63 images. The increase in predictive power of the spatial bias achieved through incorporating subject-specific information might appear small, but it is a sizable fraction of the dynamic range between lower and upper limit (0.073 and 0.173 naturals and urbans respectively), and significant for all numbers of training images (paired T-tests over 48 subjects, $p < 0.001$).

Combining the positive effects of knowing the correct subject and knowing many subjects We have seen that the prediction of the spatial bias from one independent subject can be improved on in two ways. By incorporating information from more independent subjects (see Estimating the lower bound for fixation selection models), reducing the uncertainty in the estimation of the true spatial bias, or by using subject-specific information (see Subject-specific spatial bias). Both improvements have effects of similar sizes. It seems possible that combining both methods would allow an even better prediction. We hypothesize that the spatial bias of a large set of subjects consists of certain identifiable components, to which individual subjects contribute with different strengths. In that case, it should be possible to express an individual subject's spatial bias as a combination of these components. Such an approach would be more reliable, because the components can be estimated from many different subjects, effectively reducing the noise in the estimate. To identify these components, we compute the spatial bias for all training subjects on a given number of images, and perform a principal components analysis (PCA) on these biases.

Figure 2.4 A,B shows the first 12 principal components of an exemplary case, which are the directions where the spatial bias varies most over subjects. Importantly, the amount of variance explained by the components drops rapidly (see Figure 2.4C). Hence the first few components explain the larger part of variance of the data and the remainder is increasingly noisy and uninformative. To enhance

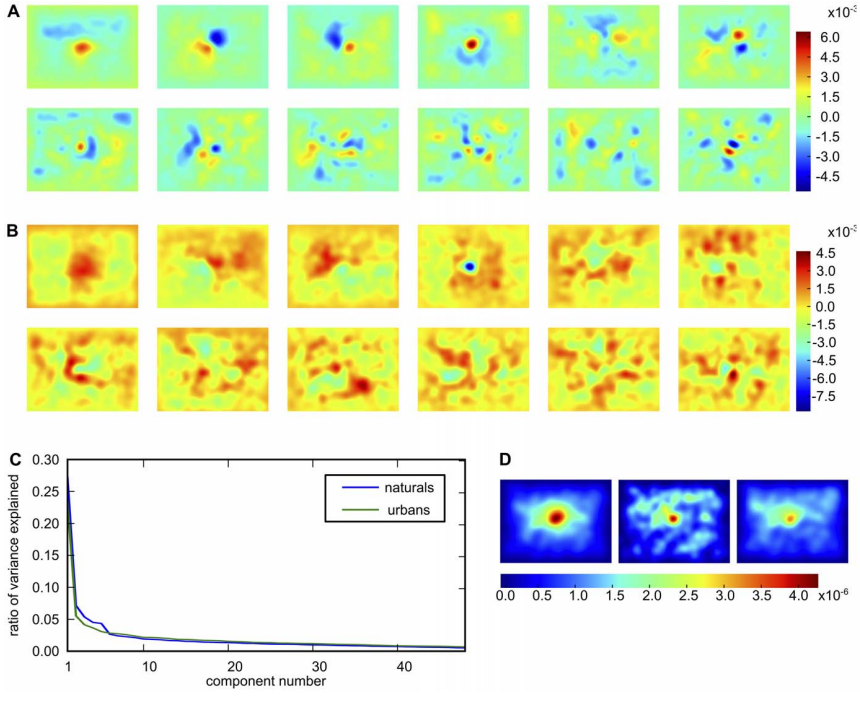


Figure 2.4: **P**CA-based cleaning of a subject-specific spatial bias. Panels A and B show the first 12 principal components respectively for naturals and urbans. For demonstration purposes, the underlying subject biases were computed with fixation data from all images and all subjects. Please note that the sign of the principal components is arbitrary. Panel C shows that the variance explained by each component drops dramatically. This, and the fact that the first 5 components carry some interpretable meaning, led us to choose the first five components for the cleaning of the subject-specific bias. Panel D shows an example of this. The left plot shows the spatial bias of all other subjects, the center one the subject-specific bias and the right plot shows the result of reconstructing the subject-specific bias with the first five principal components.

the reliability of the estimate, we only keep the first 5 components. We incorporate subject-specific traits by finding subject-individual weights for the components. These weights are computed by regressing the components onto the subject-specific bias, which is computed on all images in the training set. Figure 2.4D illustrates the subject-specific weighting of the multi-subject spatial bias. This combines the subject specific information and the statistical reliability of a large data base.

Importantly, we do not use the subject or image to be predicted for estimating the components. To evaluate the efficacy of this approach, we carry out the same subject evaluation as for the evaluation of the Subject-specific spatial bias, but use the described PCA method instead of the regular individual subject bias. This procedure combines two possible sources of improvements: subject-specific information and noise reduction in the spatial bias estimate. To ensure that the subject-specific weighting of principal components has a separate effect, we also evaluate how the PCA spatial bias cleaning without subject-specific weighting performs. For this control, we simply weight the first five components with their eigenvalues and use their sum as the prediction. In order to evaluate whether this method is able to combine the positive effects of knowing a specific subject and of having a robust estimate from many subjects, we need to compare it to both individual methods.

First we investigate the improvement in predictive power in comparison to the subject specific spatial bias (Figure 2.5). In case a single natural image is used to compute the principal components no improvement is observed. For an intermediate number of images a significant improvement (paired t-test, 48 subjects, significance level indicated by number of asterisks) compared to the subject specific spatial bias is demonstrated (Figure 2.5B upper row, significant deviation of blue dots from the horizontal axis that was the main diagonal in the original scatter plot). Testing subjects on even larger numbers of natural images leads to a smooth distribution of the spatial bias and no further improvement by PCA-cleaning is achieved.

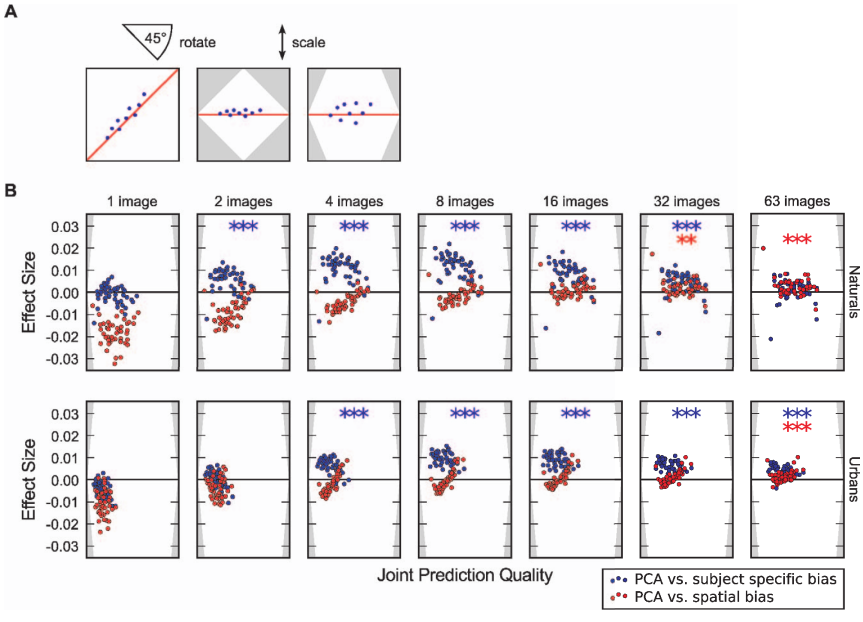


Figure 2.5: **T**he effect of using a subject-specific PCA cleaned bias for prediction. Panel A explains how the plots in B come about. We scatter the AUC score for predicting individual subjects averaged over images and repetitions with the PCA-cleaned bias against either the scores for the subject-specific or average spatial bias. For better visibility we rotate the plot by 45° degrees and sclae it. This causes the x-axis to become a measure of how well a subject can be predicted with either method and the y-axis becomes a measure of effect size, i.e. how much the prediction improves by application of the PCA. Please note, the y-axis is labeled such that it indicates the difference between the two scores and not the distance to the diagonal. To make the effects more visible we scale the y-axis to include the relevant range. The blue dots compare the effect of using PCA-cleaning to the subject-specific bias. It can be seen that in both categories the effect of the PCA depends on the number of images. The asterisks indicate that the effect size is significantly larger than zero (paired t-test, * $\widehat{=} p < 0.05$, ** $\widehat{=} p < 0.01$, *** $\widehat{=} p < 0.001$).

In the case of urban images an improvement is observed in a range from 4 to 63 images (Figure 2.5B lower row), which is shifted by a factor of two compared to naturals. Hence, in comparison to the subject specific spatial bias PCA-cleaning boosts performance by a modest degree for the case of testing with an intermediate number of images. Second, we compare prediction performance of PCA-cleaned individual spatial bias to the average obtained by a large number of subjects. Here we observe a small but significant improvement only for a larger number of images (Figure 2.5B significant deviation of red dots from the horizontal axis). The small effect size might be expected because there is already so little noise in the spatial bias for one subject. Thus, the predictive power of the generic spatial bias is already very high, leaving little room for improvement. On the other hand, the results for a small number of images illustrate that the PCA cleaning requires a certain amount of data to work properly. There is a possibility that the subject specific weights do not contribute to the observed effect, but that PCA-cleaning is only effective by removing noisy components. To control for this we repeated the same analysis but omitted the subject-specific weighting and instead weighted the components with their eigenvalues obtained from the PCA. This does not lead to a change in predictive power compared to the pure spatial bias (paired T-test, $p > 0.2$; data not shown). In summary, the PCA cleaned subject-specific spatial bias estimate combines the positive effects of reliable bias estimation and exploiting subject-specific traits.

Predicting better than perfect: combining subject- and image-specific biases
The previous section showed that subject-specific predictions can improve the already good prediction of a large group of subjects in the domain of the spatial bias. After estimating the upper bound for fixation selection models, we established that the inter-subject consistency marks an upper bound for prediction quality of subject independent models. Given these observations, the question arises whether subject-specific models can surpass the inter-subject

consistency bound. As a proof of concept, we combine inter-subject predictions with the subject-specific spatial bias as a simple form of subject-specific information, and analyze if this procedure can lead to a better prediction. We assume that viewing behavior on an image is driven partly by a subject-specific spatial bias and by image properties, i.e. the inter-subject prediction contains both components. The idea is to replace the general spatial bias in the inter-subject fixation density map with a subject-specific spatial bias while keeping the image dependent part. To achieve this, we first compute the fixation density map of all training subjects on the image in question, i.e. the inter-subject prediction. Second, we remove the general spatial bias by dividing the inter-subject prediction point-wise through the training subjects' spatial bias computed on all other images. To arrive at a prediction, we multiply the resulting image-specific bias point-wise with the spatial bias of the predicted subject. Finally, we normalize the resulting map to unit mass and evaluate how well it predicts the fixations of our test subject. We use the same cross-validation procedure as for the generic inter-subject predictions, but limit the computations to the logarithmically increasing training set sizes used for the spatial bias evaluation. Inter-subject consistency is recomputed for these new training sets to allow paired tests between subject-specific and generic predictions. The results show a small but significant effect on naturals ($p < 0.001$, paired t-test for 4 or more subjects.

See Figure 2.6). For example, the improvement for 47 subjects is a mean AUC increase from 0.809 to 0.815. There is no significant effect on urbans (paired t-test, $p > 0.2$ for all numbers of subjects). The difference between categories can probably be explained by the fact that the spatial bias has less predictive power for urbans and that the inter-subject consistency is already higher in urbans. We conclude that the combination of subject-specific information and image-specific information can surpass the inter-subject consistency upper bound on natural but not on urban images.

We draw five different conclusions: First, the lower bound, based

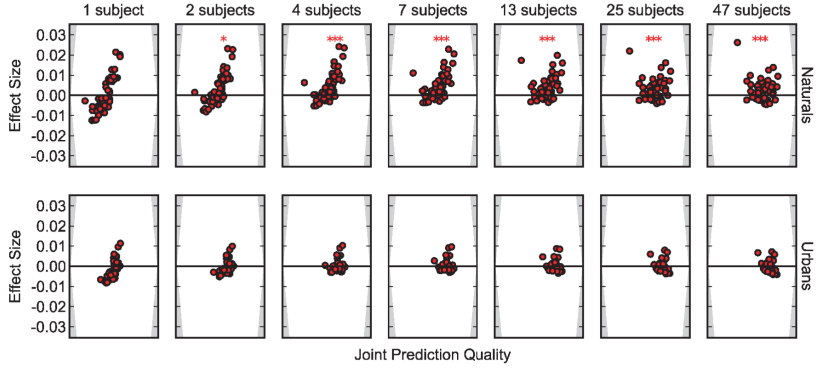


Figure 2.6: **C**ombining a subject-specific and image-specific spatial bias for a better than perfect prediction. The plots are produced as in Figure 2.5. The effect depends on the number of subjects that enter the bias estimation and the image category. For natural scenes, a statically significant effect (paired t-test, *** = $p < 0.001$) can be seen when four subjects or more are used. The effect cannot be seen for urban scenes, which might be explained by the low predictive power of the subject-specific bias compared to the high predictive power of the image-specific bias on urbans.

on the image- and subject-independent spatial bias, is surprisingly high (AUC of 0.729 and 0.673 for naturals and urbans respectively) but the reliability of the estimated bound depends on the size of the data set. For all practical purposes, 32 images and 25 subjects seem to be sufficient for a reliable estimate. Second, the reliability of the upper bound, which is based on the consistency of viewing behavior between subjects, also depends on the data set size. For all practical purposes, the inter-subject consistency of about 20 subjects is sufficient to establish an upper bound for generic models of fixation selection in free viewing tasks. Third, the incorporation of subject-specific information can significantly improve the predictive power of the subject- and image-independent spatial bias. Fourth, the predictive power of the spatial bias can further increase when the subject-specific information is de-noised with information from other subjects. Fifth, the dependence of the upper bound on joint-subject processes makes it possible to surpass this bound by combining subject- and image-specific biases.

2.4 Discussion

In this work, we have focused on how models of fixation selection can be evaluated. Based on theoretical considerations, we argued that the AUC is the best choice for the kind of data that is usually available in eye-tracking studies. However, when predicting viewing behavior that is consistent across a group of subjects, KL-divergence presents itself as a superior alternative, given that the data set is large enough. Regardless of the measure, model evaluation is also influenced by the inherent properties of eye-tracking data. In particular, the predictive power of the pure spatial bias estimate poses a challenging lower bound for prediction performance that any useful model has to exceed. Moreover, the inter-subject consistency constitutes an upper bound for generic models of fixation selection. The accuracy of the estimate for both bounds depends decisively on data set size.

By using these bounds as a reference frame, we showed that sub-

ject idiosyncrasies can be exploited to increase the prediction performance. This can be pushed to the point where the predictive power surpasses the inter-subject consistency bound. From a more general perspective, the two bounds discussed in this paper form a reference frame that allows for a substantially more informed assessment of the quality of a model of fixation selection than just a measure score alone. It is essential that these bounds are reliably estimated by acquiring enough data. To see this, consider a case in which data is only available from a small set of 10 subjects. In this case the inter-subject AUC and the predictive power of the spatial bias will be underestimated. Both these effects subsequently lead to an overestimation of model quality. The following two examples illustrate the advantages of our approach when this caveat is kept in mind.

First, if we consider a task that induces a very specific spatial bias (e.g. pedestrian search, Torralba et al., 2006), the AUC score depends on how much of the image is covered by the task-relevant area. People will look for pedestrians on the ground, so in principle it is possible to increase the area of the sky, e.g. by decreasing the camera's focal length, without substantially changing fixations patterns. If our model has also learned to ignore that additional spatial region, the AUC is increased substantially. Yet we would not claim that the increased AUC reflects a better description of the fixation selection process. Reporting the predictive power of the pure spatial bias alongside the model's score allows a fair evaluation of a model in all cases.

Secondly, in our data we found that the category where the spatial bias is weaker (urbans) has a stronger inter-subject consistency. This double-dissociation has important consequences for the evaluation of fixation selection models. One and the same model, incorporating both spatial bias and image statistics, may score higher on naturals than on urbans, because of the predictive power of the spatial bias. On the other hand, if a model is almost optimal and comes close to the predictive power of other subjects' fixations, it will score higher

on urbans. Thus, the type of dataset the model is evaluated on will have an effect on one's judgment of model quality. As a result of this, a comparison of different models is nearly impossible if they were evaluated on different data sets, unless the upper and lower bounds for the specific datasets are explicitly given.

A different, commonly used method to control for the spatial bias when using AUC is to sample the negative observations not from the whole image, but only from points that have been fixated on other images (Tatler et al., 2005). If this is accompanied by an equally corrected report of inter-subject consistency, it allows for an unbiased model comparison much in the same way as reporting upper and lower bounds as proposed here. In the context of model evaluation, however, we believe that explicit is better than implicit, i.e. that reporting the complete reference frame gives the reader a more direct grasp of the model's capabilities. We conclude that the most comprehensive way to evaluate a model of fixation selection, especially with respect to comparisons between different models, is to use AUC and/or KL-divergence as performance measures, and to report both the predictive power of the spatial bias and the inter-subject consistency of the data set that the model is tested on.

Besides putting model performance into perspective, the proposed reference frame can also be of use prior to model evaluation. The two bounds define the dynamic range for predictions of the distribution of fixation points. The ideal data set for evaluating a model of fixation selection would have a large range, indicating that subjects fixate different locations on different images - limiting the predictive power of the spatial bias - but agree on the selection of fixation points on single images. When the predictive power of the spatial bias is small, models of fixation selection can only improve by uncovering regularities distinct from the spatial bias. At the same time, high inter-subject consistency indicates that a common process regulates the selection of fixations in observers, and it is this process that models of fixation selection target.

With a change in perspective, the reference frame can be used to

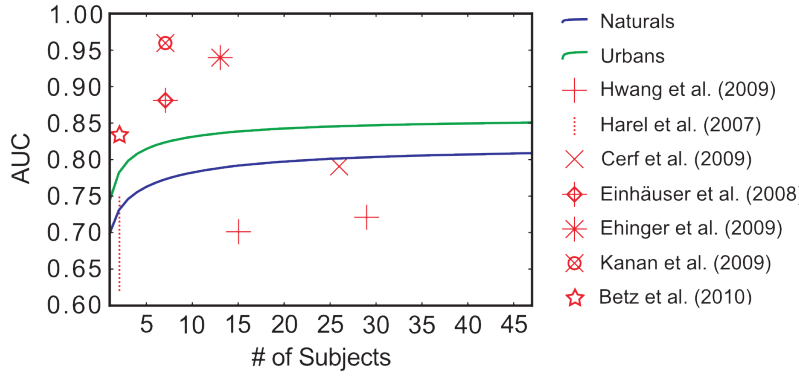


Figure 2.7: **A** comparison of inter-subject consistency AUC in different studies. Green and blue lines show the dependence of inter subject consistency on the number of subjects in our data. The symbols show inter-subject consistency values reported in other studies. All studies that reported higher values used either stimuli that contained a wealth of high-level information or employed a specific task. Cerf, Frady, and Koch (2009) also use a free viewing task and are compatible with our findings. Harel, Koch, and Perona (2007) only report a range of values (read from a figure). Notably, Hwang, Higgins, and Pomplun (2009) use image rotations to diminish top-down influences and observe lower inter-subject consistency.

probe for differences in viewing behavior. The lower bound indicates to what extent subjects' viewing behavior is independent of the image, whereas the upper bound quantifies their agreement. This not only allows interesting comparisons between different groups of subjects, but also provides a tool to investigate the effect of different stimulus categories. In this work, we investigated urban and natural images and found that the range of the reference frame is larger on urban than on natural images. This shows that urban images elicit higher subject agreement in fixation selection and evoke a stronger image-dependent component in fixation target selection. The cause of the differences between categories is an interesting topic for further investigation.

The inter-subject consistency has been used before as an upper

bound for model performance, which allows for a direct comparison of our values and the ones provided in the literature. Interestingly, we found that on first sight not all values were in line with our results (Figure 2.7). However, there seems to be a consistent explanation for the deviations: All values of inter-subject consistency that lie above those found in our data were computed on data where there was an explicit task during the eye-tracking experiment (object naming (Einhäuser et al., 2008) or pedestrian search (Ehinger et al., 2009; Kanan et al., 2009)), or the stimuli material contained a wealth of high-level information (web-pages (Betz et al., 2010)). On the other hand, Hwang et al. (2009) explicitly designed their experiment to minimize high-level information by rotating the images by 90° or 180°. They report lower inter-subject consistency, but the effect of group size is in line with our results. Finally, Cerf et al. (2009) use a free viewing task similar to our experiment and obtain values almost identical to ours. We conjecture that inter-subject consistency is strongly influenced by the subjects' task and the availability of high-level information. This is also in line with the category differences found in our data (urbans $>$ naturals), since the urban scenes provide more high-level information (e.g. man-made objects, people), as well as with category differences reported by Frey et al. (2008). Interestingly, high inter-observer consistency is not related to a large influence of the spatial bias. In our dataset, the former is higher for urban scenes while the latter is higher on natural images. A speculative explanation of this finding is that when high-level information is present in an image, it will guide the eye movements of many subjects to locations that are not necessarily in the center of the image, increasing inter-subject consistency and decreasing the influence of the spatial bias. In the absence of high-level information, subjects tend to look more towards the center of the screen, but in a less homogenous fashion. This fallback strategy leads to an increased spatial bias and decreased inter-subject consistency. Further evidence for this hypothesis comes from eye-tracking studies with pink-noise stimuli, which are completely devoid of high-level

information and where the influence of the spatial bias is comparatively large (Açik, Sarwary, Schultze-Kraft, Onat, & König, 2010). Our analyses of subject idiosyncrasies relative to our established bounds showed that the increase in performance, although statistically significant, is very small. In the case where data from 63 images are used, knowing the spatial bias of a specific subject is as good as knowing more than 7 other subjects on naturals, or knowing more than 2 other subjects on urbans. The smaller effect for urbans fits the observation that inter-subject consistency is higher in that category, making knowledge about a specific subject less unique. This relates to a possible reason for the small overall effect size in both categories: Açik et al. (2010) show that different demographic subject groups have remarkably different viewing behavior. Specifically, explorativeness, a property that is closely related to the spatial bias, decreases with increasing age. Our subject group consisted exclusively of university students between 19 and 28 years of age. Thus it can be expected that the effect of knowing the subject to be predicted would be much larger in a more heterogeneous subject group with lower inter-subject consistency. In such a scenario, the improvement caused by PCA-cleaning demonstrated in the present study could become more relevant. In general, the PCA-cleaning requires fixation data on a fair number of images for a good signal to noise ratio. In practice, the principal components could be determined from a large set of subjects and images recorded in a baseline study. It may then be possible to tailor a clean subject-specific spatial bias based on fixations from the subject of interest on few images. This technique may be useful in a modeling context, when the goal is to fine-tune a generic model for predicting individual subjects' fixations.

The spatial bias is of course only one feature of viewing behavior where subject idiosyncrasies can play a role. There are possibly many different ways to incorporate these into a model of fixation selection. An obvious candidate would be the relative importance of different image features in a bottom-up model. Whether subject-specific mod-

eling of feature weights has a positive effect is an interesting question for further research, but goes beyond the scope of this article.

Finally, we showed that it is possible to surpass the limit set by the inter-subject consistency when incorporating subject and image-specific information into the prediction. Despite the very small effect, this result exemplifies the potential value of subject-specific predictions. However, it also reveals another aspect of the evaluation of models of fixation selection. Judging only by the AUC values, we have created a prediction that exceeds the inter-subject consistency bound and incidentally also the best prediction ever described in the literature. In a sense, our prediction is better than what has previously been called 'perfect'. Of course no sensible person would congratulate us on this achievement. Rather, it shows that claims about theories of fixation selection based purely on a prediction's AUC values, or the percentage of inter-subject AUC achieved, can be quite hollow.

A decisive question that should be part of every model evaluation is what we can learn from this model about processes of fixation selection implemented in the brain. Good models do not only achieve high prediction scores, but also reproduce and, better, explain differences in human viewing behavior, such as the different reference frames between natural and urban images, or the temporal evolution of scan paths. Models that replicate novel aspects of viewing behavior might still be revealing about the underlying mechanisms, despite having low predictive power. Here, we have to consider two questions: do we understand the mechanism by which our model goes from input to prediction? And is this mechanism plausible? If we can answer both these questions in the affirmative, and our model performs well on an adequate stimulus set under the evaluation procedures described in this article, we really will have made a contribution.

2.5 Materials and Methods

Theoretical maximum value for AUC

In the present work, receiver-operating characterisics (ROC) (see Existing measures) analysis is applied to classify fixated locations vs. non-fixated locations. This treats the prediction of fixations as a discrete binary problem: a location is either fixated or it is not. However, for an unbounded number of subjects and taking into account finite precision of the occulomotor system and the eye-tracker, there is no principled reason why a location cannot be fixated and therefore all locations should eventually be fixated. This implies that every location has a finite probability to be selected as fixated and a finite probability to be selected as non-fixated. Hence, classification of a location inherently carries an error, as it is neither perfectly fixated nor non-fixated. It follows that an AUC of 1 is not achievable and a bound lower than 1 does exist.

In this section, we formalize these considerations and derive a quantitative estimate of the upper bound of the area under the ROC curve when we conceptualize the prediction as a probability density function. In the following we redefine the hit and false alarm rate for calculating the AUC value to work with probability distributions. The observed distribution of fixation points upon presentation of stimulus i is described by $efm_i(x)$, with $0 \leq efm_i(x) \leq 1$ for all $x = 1 \ldots n$. The 2D topology is irrelevant, as there is no interaction between different positions, hence we can use a one dimensional index. Furthermore $\sum_x efm_i(x) = 1$. We assume that for all x $\sum_i efm_i(x) = const$. This means that for every location, across all images, the probability of fixations is constant, i.e. there is no spatial bias. A spatial bias leads to additional complications like equilibrating the spatial discretization to achieve a constant distribution of control (non-fixated) locations. It does, however, not change the principle result. We furthermore assume that the prediction of fixated regions $pfm(x)$ is perfect when $pfm(x) = efm(x)$. Now we evaluate the quality of this prediction in terms of ROC. For a

threshold θ the number of hits is given by

$$hit(\theta) = \sum_{\forall x \in \{pfm(x) > \theta\}} (efm(x))$$

We classify as a fixated all locations where the prediction exceeds the threshold, and weight each such location with the empirical probability that this point is fixated. Above we assumed pfm equals efm and we simplify

$$hit(\theta) = \sum_{\forall x \in \{efm(x) > \theta\}} (efm(x))$$

Because of all x $\sum_i efm_i(x) = const$ and $\sum_x efm_i(x) = 1$ the distribution of control fixations is flat at a value of $\frac{1}{n}$ and the number of false alarms is

$$fa(\theta) = \sum_{\forall x \in \{pfm(x) > \theta\}} (1/n)$$

Again we count all locations where the prediction exceeds the threshold, but now weight each such location with $\frac{1}{n}$. As before, the predicted map equals the empirical one and we have

$$fa(\theta) = \sum_{\forall x \in \{efm(x) > \theta\}} (1/n)$$

For any non-degenerate distribution where efm takes on values other than 0 and 1 there must be a threshold where $hit(\theta) < 1$ and $fa(\theta) > 0$. Hence the area under the ROC curve is smaller than 1.

What is the upper boundary of the AUC for a specific efm? Given

$$hist : efm(x) -> h(s),$$

with $h(s)$ the frequency of occurrence of a specific saliency value s. $h(s)$ has some important properties:

$$\int_0^1 h(s)\,ds = n$$

the spatial discretization of $efm(x)$ is n and because $\int_x efm(x) = 1$ also

$$\int_0^1 h(s) \cdot s\,ds = 1$$

is a probability density distribution with integral 1. For a given θ the false alarm rate is given by

$$fa(\theta) = 1/n \int_{s=\theta}^1 h(s)\,ds$$

The integral yields the number of points above the threshold which is weighted with $1/n$. The hits are given by

$$hit(\theta) = \int_{s=\theta}^1 h(s) \cdot s\,ds$$

When using these definitions of hits and false alarms the AUC is given by

$$AUC(h) = \int_{fa=0}^{fa=1} hit(fa)\,dfa$$

Note that the false alarm rate increases as we lower the threshold from 1 downward. By change of variables we obtain

$$AUC(h) = \int_{\theta=1}^{\theta=0} hit(fa)\frac{dfa}{d\theta}\,d\theta$$

changing the bounds

$$AUC(h) = \int_{\theta=0}^{\theta=1} (-1) \cdot hit(fa)\frac{dfa}{d\theta}\,d\theta$$

As $\frac{dfa(\theta)}{d\theta} = -h(s)$ (see definition of fa above) we obtain

$$AUC(h) = \int_{\theta=0}^{\theta=1} (-1) \cdot hit(fa)(-1) \cdot h(s)\, d\theta$$

$$AUC(h) = \int_{\theta=0}^{\theta=1} hit(fa(\theta)) \cdot h(\theta)\, d\theta$$

$$AUC(h) = \int_{\theta=0}^{\theta=1} \int_{s=\theta}^{s=1} h(s) \cdot s\, ds\, h(\theta)\, d\theta$$

This formula yields the upper bound for predicting a given empirical fixation map.

Proof of AUC linearity

Here, we prove that the value of the area under the receiver-operating characteristics curve (AUC) for a given multiset of positive (P) and negative (N) observations does not depend on how the positive observations are grouped, i.e.

$$AUC(P_1 \uplus P_2, N) = \frac{|P_1|}{|P_1 \uplus P_2|} \cdot AUC(P_1, N) + \frac{|P_2|}{|P_1 \uplus P_2|} \cdot AUC(P_2, N) \tag{2.1}$$

where $\uplus$ denotes the multiset union. As a given location may be fixated several times the notion of a multiset seems appropriate. Multisets are a generalization of sets and may contain multiple memberships of one and the same element. The AUC is obtained through trapezoidal approximation of the area under the curve plotting the true positive rate (TPR) against the false positive rate (FPR) for all thresholds, according to:

$$AUC(P, N) = \sum_{i=2}^{n} \frac{TPR(t_i) + TPR(t_{i-1})}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \tag{2.2}$$

$$TPR(t) = \frac{|\{x|x \in P \wedge x \geq t\}|}{|P|} \tag{2.3}$$

$$FPR(t) = \frac{|\{x|x \in N \wedge x \geq t\}|}{|N|} \tag{2.4}$$

$$t_1 = \infty,\ i < k \Rightarrow t_i > t_k,\ t_n = -\infty \tag{2.5}$$

Lemma. Let $S \in \mathcal{P}(\mathbb{R})$ be a finite set of real numbers and $f : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ be a function, such that for each $m \in S$ hold

$$f(S) \cdot |S| = f(S \setminus \{m\})(|S| - 1) + f(\{m\})$$

That implies for any set $T \subseteq S$

$$f(S) \cdot |S| = f(S \setminus T)(|S| - |T|) + f(T) \cdot |T| = \sum_{s \in S} f(\{s\}).$$

This can easily be seen through induction over $|T|$, beginning by $T = \emptyset$

The Lemma reduces (2.1) to

$$AUC(P, N) = \frac{|P| - 1}{|P|} \cdot AUC(P \setminus \{p\}, N) + \frac{1}{|P|} \cdot AUC(\{p\}, N) \tag{2.6}$$

From (2.3) follows

$$\forall p \in P, \quad TPR_{P \setminus \{p\}}(t) = \begin{cases} \frac{TPR_P(t) \cdot |P| - 1}{|P| - 1} & \text{if } t \leq p \\ \frac{TPR_P(t) \cdot |P|}{|P| - 1} & \text{if } t > p \end{cases} \tag{2.7}$$

Now we can compute $AUC(P \setminus \{p\}, N)$ and $AUC(\{p\}, N)$. Let $k \in [1, n]$ be the smallest value for which $t_k > p$, then

$$
\begin{aligned}
&AUC(P \setminus \{p\}, N) = \\
&\sum_{i=2}^{k} \frac{(TPR_P(t_i) + TPR_P(t_{i-1})) \cdot |P|}{2 \cdot (|P| - 1)} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad + \sum_{i=k+1}^{n} \frac{(TPR_P(t_i) + TPR_P(t_{i-1})) \cdot |P| - 2}{2 \cdot (|P| - 1)} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&= \frac{|P|}{|P| - 1} \cdot \sum_{i=2}^{k} \frac{TPR_P(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad + \frac{|P|}{|P| - 1} \cdot \sum_{i=k+1}^{n} \frac{TPR_P(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad - \frac{1}{|P| - 1} \cdot \sum_{i=k+1}^{n} FPR(t_i) - FPR(t_{i-1}) \\
&= \frac{|P|}{|P| - 1} \cdot AUC(P, N) - \frac{FPR(t_n) - FPR(t_k)}{|P| - 1} \\
&= \frac{|P|}{|P| - 1} \cdot AUC(P, N) - \frac{1 - FPR(t_k)}{|P| - 1}
\end{aligned}
\tag{2.8}
$$

and

$$
\begin{aligned}
AUC(\{p\}, N) &= \sum_{i=2}^{n} \frac{TPR(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&= \sum_{i=2}^{k} 0 + \sum_{i=k}^{n} \frac{1 + 1}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&= FPR(t_n) - FPR(t_k) \\
&= 1 - FPR(t_k)
\end{aligned}
\tag{2.9}
$$

Using (2.8) and (2.9) it is easy to see that (2.6) is true, proving

(2.1).

Computational details of AUC analysis

Although in theory AUC is independent of arbitrary parameters, this is not entirely true in practice. Strictly speaking,the ROC curve plots the probability of a hit against the probability of a false alarm, and these probabilities of course have to be estimated. However, we have found that when applying this measure to the evaluation of models of fixation selection, using relative frequencies as an estimation of probabilities works well and can be seen as a sensible default value that requires no further parameters. In that case, there remain two decisions on related issues that have to be made when computing the AUC, and both influence the resulting value: first, we need to decide which thresholds to use to create the underlying ROC curve, since an infinite number of thresholds with infinitesimal spacing is not achievable. Second, it has to be decided how the area under the ROC curve is computed. In general, trapezoidal integration is the method of choice. However, in the special case of fixation classification, there is a simpler way. Here, it is usually the case that we have a very large number of negative values (either all values in the salience map, or all values that were not fixated, or all values at locations that were fixated on other images) and a smaller set of positive values (salience values at fixated locations). Obviously it suffices to use all unique values in the combined set of positives and negatives as thresholds. Neither the true positive rate nor the false positive rate will change for any other threshold values. In general, the true positive rate can only increase for threshold values in the set of positives. All other thresholds, those in the set of negatives, can only increase the false positive rate while the true positive rate remains constant. This implies that the ROC curve approaches a step function and the thresholds in the set of actuals define the steps. In a step function, there is no difference between trapezoidal integration and lower sum integration. And since the thresholds from the set of actuals define the steps, it suffices to use lower sum integration with only these

values as thresholds. There is one pitfall that has to be avoided with this approach. When no threshold reaches a true positive rate of one before the false positive rate is one, the AUC can be underestimated. If this is the case, we use trapezoidal integration for the last segment of the curve. This method, which is computationally much more efficient, as it involves fewer threshold values, was adopted for all reported AUC values in this article.

Fixation density map estimation

In the analysis of eye-tracking data, we make frequent use of fixation density maps (FDM), which estimate the probability that a specific location is fixated. These are computed by smoothing a two-dimensional histogram of fixations, where each pixel is one bin, with a Gaussian kernel of 2° FWHM, normalizing to unit mass. The rationale for smoothing is that a) the eye-tracker operates with limited resolution (calibration-error $< .3°$) and b) the visual system samples information at high-resolution not only from a single fixated pixel but from the fovea which corresponds to about 2° of visual angle in diameter. For computational efficiency it is often necessary to scale FDMs to smaller size. This is achieved by adjusting the bin sizes of the histogram and the size of the Gaussian kernel accordingly.

Correction of KL divergence for small samples

The KL-divergence can be expressed in terms of Information Entropy, and for Information Entropy it is known that it systematically depends on the sample size (Hausser & Strimmer, 2009; Miller, 1955; Nemenman, Shafee, & Bialek, 2002). These observations lead us to suspect that the KL-divergence is also biased, which is problematic when different models are evaluated against densities estimated from different sample sizes. We carry out two simulations to investigate the size of this potential confound. First, we treat the overall spatial bias as our prediction. We then take a random sample of fixations from the set that constitutes the spatial bias and repeatedly

calculate the KL-divergence between the FDM of our sample and our prediction. If the sample gives a perfect estimate of the distribution it was drawn from the KL-divergence should be zero. We increase the number of fixations per sample from 6 to 800 in steps of 2, and draw 1000 samples of every size. Since discrete Entropy estimates are also strongly influenced by the binning of the probability density function, we do not use our standard procedure for computing fixation density maps. Instead, we sort the data into a grid of 16x12 bins (leading to N=192). The number of grid cells was selected such that the area of each bin is equal to the area of a circle of diameter two degrees of visual angle. These FDMs are not smoothed, since they already have a coarse resolution. In a second simulation, we take a normal distribution with specified parameters ($\mu = 0$, $\sigma = 1$) as our prediction and sample our data from a different normal distribution ($\mu = 2$, $\sigma = 1$). In this case the true KL-divergence can be determined analytically and the KL-divergence computed from different sample sizes can be compared to this target value. We proceed in the same way as before and increase the sample size from 6 to 800 in steps of 2 and draw 1000 samples of every size. Densities are estimated as histograms with 100 bins.

In both cases the estimated KL-divergence was higher than the analytical value. The difference between mean estimated KL-value and analytical value decreased with increasing sample size (the results for simulation 1 are depicted in Figure 2.8A; results for simulation 2 were similar). Thus, comparing models evaluated on different data set sizes is difficult. One approach to cope with the sample size dependence of the estimate is to keep the sample size constant in every comparison by randomly sampling as many fixations from each data set as are available from the smallest one. However, if the size of a novel data set is comparably small and previous model evaluations were performed on a larger and inaccessible data set, it is not possible to reduce the larger data set. Thus, to foster comparisons between different studies, it would be advantageous to be able to directly correct for the bias introduced by sample size.

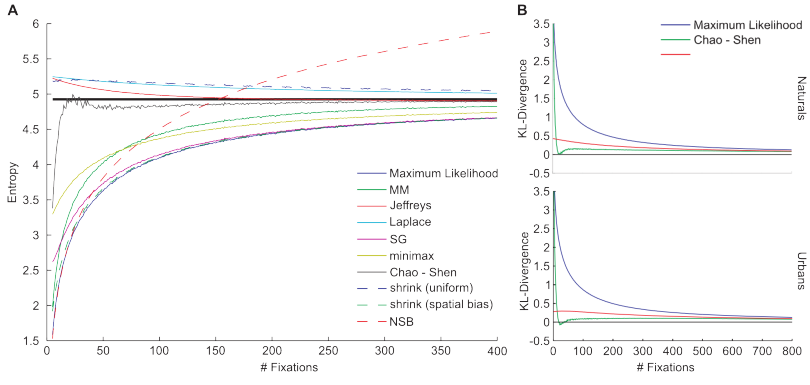


Figure 2.8: **T**he effect of sample size on the KL-divergence. A. Performance of different methods to remove the sample size bias from entropy estimates in a simulation using eye-tracking data. The bold line shows the maximum likelihood entropy estimate computed on the entire data set ($N > 40000$)and can be interpreted as ground truth. The Chao-Shen and Jeffreys correction methods approach the target value with the lowest number of samples. Descriptions of the individual methods can be found in (Chao-Shen: Chao & Shen, 2003), (shrinkg: Hausser & Strimmer, 2009), (Laplace Holste, Grosse, & Herzel, 1998), (Jeffreys Krichevsky & Trofimov, 2002), (MM Miller, 1955), (NSB Nemenman, Shafee, & Bialek, 2002), (SG Schürmann & Grassberger, 1996), (minimax Trybula, 1958). B. Sample size dependence of different KL-divergence estimation methods. The standard maximum likelihood method shows a strong positive bias for small samples, both correction methods tested can reduce this problem for sample sizes of ca. half the number of bins in the estimated distributions or larger.

There are multiple methods that try to improve the estimate of entropy values (recall that KL-divergence is directly dependent on the Entropy estimates), as compared to the typically-used maximum likelihood approach. We therefore investigate the applicability to fixation data of several methods (Chao & Shen, 2003; Hausser & Strimmer, 2009; Holste, Grosse, & Herzel, 1998; Krichevsky & Trofimov, 2002; Miller, 1955; Nemenman et al., 2002; Schürmann & Grassberger, 1996; Trybula, 1958), for which Hausser and Strimmer (2009) provides an implementation. To compare the efficacy of the different approaches, we carried out simulations in which we estimated the entropy of differently sized samples from the general spatial bias. In addition to the direct relevance for the calculation of KL-divergence, an important advantage of an unbiased entropy estimate is that entropy can be used to characterize viewing behavior (Açik et al., 2010; Gilland, 2008; Recarte & Nunes, 2000). It is therefore relevant to have an unbiased estimate, e.g. for comparing different experimental conditions with different amount of fixations. The simulations follow the pattern that we used for determining the sample size dependence in KL-divergence. Due to the large number of different correction methods compared, we only draw 200 samples of each size to reduce computational load. We compare estimates for different sample sizes to the entropy of all fixations in one category ($N_{naturals} = 43295$,$N_{urbans} = 44753$), assuming that the estimate is nearly unbiased with such a large sample size. The simulations show that it is in principle possible to improve the entropy estimate. However even in the best case, the number of samples required for a reasonable estimate is approximately half the number of bins of the fixation density map. This is a large improvement over uncorrected Entropy, which requires the number of data points to be at least equal to the number of bins squared. The fixation densities in our simulations were down sampled to 169 bins. Considering that FDMs are typically smoothed with a 2deg FWHM Gaussian kernel, the effective resolution of a FDM is already much lower than the number of pixels suggests, making the down sampling tenable.

Overall the correction methods proposed by Chao and Shen (2003) and Krichevsky and Trofimov (2002) work best of all tested methods. To yield a correction method for the KL-divergence, its Entropy and cross-Entropy terms have to be corrected. Starting with Chao-Shen, the pure entropy term can straightforwardly be corrected. Moreover, if we presuppose that a model output corresponds to a correct probability density (Q), we can also apply Chao-Shen to correct the cross Entropy $H(P, Q)$. Here, we use

$$H(P) = -\sum_i \frac{p_i^{cs} * log(p_i^{cs})}{Coverage(p_i^{cs})}$$

$$H(P||Q) = -\sum_i \frac{p_i^{cs} * log(q_i)}{Coverage(p_i^{cs})}$$

to compute the corrected KL-divergence, where pcs and Coverage are the two Chao-Shen correction terms (see Chao and Shen, 2003). The Jeffreys correction can simply be applied by adding $1/2$ to the cell counts of the FDM before it is normalized to unit mass. To validate applicability of Chao-Shen in the case of KL, we repeated the simulations for the maximum likelihood KL-divergence estimation but used the Chao Shen and Jeffreys corrected estimation. As shown in Figure 2.8B, the correction substantially improves the KL estimates as compared to the maximum likelihood version. The Jeffreys correction works well on our data, which is in part due to the fact that our distribution does not deviate too much from the uniform prior assumed by the correction method. If there are strong reasons to believe that one's data deviate much from a uniform distribution, one should therefore be careful with this correction. The Chao Shen correction is very close to the true KL-divergence between the underlying distributions at a sample size of about $N/2$.

Description of the eye-tracking study

The study has been approved by the ethics committee of the University of Osnabrück and was conducted according to the principles

expressed in the Declaration of Helsinki. All subjects gave written informed consent prior to the study and were informed of their right to withdraw at any time without negative consequences. The experiment consisted of the presentation of 255 stimuli from four different categories (naturals, urbans, fractals and pink-noise). The 'natural' category contains 64 stimuli that depict outdoor scenes like landscapes, forests and flowers. The 64 'urbans' show rural and city scenes with many man-made structures. The images comprise a large variety of different scenes and vary over many different parameters (street scenes, buildings, differences in depth and openness, close-ups and landscape perspectives). In the urban scenes only very few persons are shown and very little text. All stimuli have a large depth of field to avoid the guidance of eye movements by the photographer. We do not use the artificial stimuli from the fractal and pink-noise categories. The task of the subjects was to freely view the pictures ('watch the images carefully'). Each stimulus was shown for six seconds and a fixation point was shown in the center of the screen before each stimulus to perform a drift correction. The distance to the screen was set at 80 cm; the display used was a 21-inch CRT monitor (SyncMaster 1100 DF 2004, Samsung Electronics, Seoul, South Korea) with a screen resolution of 1280 x 960 pixels; refresh rate was 85 Hz. The stimuli had a size of approximately 28.4 x 21.3 degrees. 48 subjects (24 male) participated in the experiment and received either 5€ or course credit as compensation. Subjects were aged between 19 and 28 years, naïve to the purpose of the study and had normal or corrected-to-normal vision. The eye-tracker used was an Eyelink II system (SR Research Ltd., Mississauga, Ontario, Canada). This head-mounted system is capable of tracking both eyes; however, only the eye giving a lower validation error after calibration was used for data analysis. Sampling rate was set at 500 Hz. Saccade detection was based on three measures: eye movement of at least 0.1°, with a velocity of at least 30°/sec and an acceleration of at least 8000°/sec^2. After saccade onset, minimal saccade velocity was 25°/sec. The first 15 free fixations of each trial were used for

data analysis. All data is available from the authors upon request.

Reference values for spatial bias and inter-subject consistency

Here we report numeric AUC (Table 2.2 and 2.3) and NSS (Table 2.4 and 2.5) values for predicting fixations of one subject on one image with a subject and image independent spatial bias (estiamted lower bound, see Estimating the lower bound for fixation selection models) and with an image-specific bias (inter-subject consistency, estimated upper bound, see Estimating the upper bound for fixation selection models). All reported values are means across cross-validation runs, as described in Estimating the lower bound for fixation selection models. So far we omitted the computation of upper and lower KL-divergence boundaries. Testing the estimation reliability by changing the number of subjects and images in the training set would be confounded by the different numbers of fixations in the training set (our correction methods are intended for controlling the test set and thus do not apply here). To nevertheless be able to report sensible reference bounds, we restrict ourselves to a large training set size such that the influence of different amounts of fixations in the training set is small. In detail, we pick out one row (63 images, varying the number of subjects for prediction) and one column (25 subjects, varying the number of images for prediction) of the subject and image independent predictions. This leaves either many images or many subjects in the training set, such that there are at least 375 fixations in the training set. To furthermore minimize the effect of different amounts of fixations in the training set, we bin the screen into $12x16$ squares. The test set always contains fixations from 23 subjects, we omit the case where more than 25 subjects are in the training set, such that the number of fixations is constant at 345 fixations. The evaluation of the entropy correction methods has shown that with this amount of fixations and dimensionality of the probability density map, no correction for different amounts of fixations is needed. We also compute the inter subject consistency for predicting 23 subjects with data from the remaining 25 subjects

for every image and 48*63 random assignments of subjects into test and training set. Table 2.6 and 2.7 report the mean over images and random assignments.

Open-source python toolbox

To foster model comparison and ease reproduction of our results we provide a free open-source python toolbox. It allows to conveniently represent fixation data and can be used to estimate the lower and upper bound for fixation selection models on a given data set. Implementations of AUC and KL-divergence, as well as a few other measures, are also contained in the toolbox. The toolbox can be accessed at https://github.com/nwilming/ocupy. Furthermore, the data used in the current work is available from the authors upon request.

Acknowledgments

We thank Selim Onat, Anke Walter and Steffen Waterkamp for devising and conducting the eye-tracking experiment. We furthermore thank Hannah Knepper for carrying out the simulations for the KL-divergence correction methods. We thank Christopher Schuller for proofreading of the manuscript.

Tables

Table 2.1: Summary of described evaluation measures.

	S_a	NSS	M_{fix}/M_r	80^{th}	AUC	naïve Bayes	KL	correlation
Intuitive Scale	-	0	-	+	+	+	-	0
Few Parameters	+	+	-	-	+	-	-	-
Robustness	-	-	+	+	+	-	-	-
Low data demand	+	+	+	+	+	-	-	-

The table shows a summary of the evaluation measures and their performance with regard to the desirable properties described above. '+' indicates that the measure exhibits the property, while '0' and '-' indicate that the measure is neutral w.r.t. to the property or does not exhibit it.

Table 2.2: AUC values for natural scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47	Subject-specific
Image-specific	0.689	0.724	0.748	0.763	0.778	0.791	0.802	
63	0.703	0.715	0.723	0.726	0.727	0.728	0.729	0.732
32	0.693	0.708	0.718	0.722	0.724	0.726	0.726	0.722
16	0.678	0.696	0.709	0.715	0.719	0.721	0.722	0.707
8	0.662	0.680	0.695	0.704	0.709	0.713	0.715	0.689
4	0.647	0.661	0.677	0.688	0.696	0.701	0.704	0.674
2	0.636	0.645	0.657	0.668	0.680	0.686	0.690	0.659
1	0.619	0.631	0.643	0.651	0.660	0.668	0.675	0.640

Table 2.3: AUC values for urban scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47	Subject-specific
Image-specific	0.731	0.770	0.796	0.813	0.827	0.838	0.846	
63	0.652	0.662	0.667	0.670	0.672	0.672	0.673	0.669
32	0.639	0.652	0.659	0.663	0.665	0.667	0.667	0.657
16	0.623	0.637	0.646	0.652	0.655	0.657	0.658	0.640
8	0.608	0.619	0.630	0.636	0.640	0.643	0.645	0.624
4	0.598	0.605	0.612	0.619	0.624	0.627	0.629	0.612
2	0.593	0.596	0.601	0.604	0.609	0.610	0.612	0.603
1	0.581	0.588	0.592	0.597	0.599	0.600	0.604	0.590

Table 2.4: NSS values for natural scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47	Subject-specific
Image-specific	0.741	0.941	1.159	1.319	1.465	1.571	1.638	
63	0.773	0.835	0.871	0.887	0.897	0.903	0.905	0.976
32	0.730	0.804	0.850	0.870	0.882	0.890	0.893	0.929
16	0.664	0.752	0.810	0.837	0.855	0.865	0.870	0.854
8	0.574	0.672	0.744	0.781	0.807	0.822	0.829	0.748
4	0.472	0.570	0.653	0.699	0.732	0.753	0.764	0.623
2	0.368	0.461	0.536	0.593	0.634	0.657	0.666	0.492
1	0.277	0.346	0.439	0.490	0.520	0.559	0.577	0.376

Table 2.5: NSS values for urban scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47	Subject-specific
Image-specific	1.020	1.279	1.533	1.708	1.853	1.954	2.013	
63	0.519	0.559	0.581	0.593	0.600	0.604	0.605	0.613
32	0.470	0.519	0.549	0.564	0.572	0.578	0.581	0.559
16	0.403	0.459	0.496	0.515	0.528	0.534	0.538	0.483
8	0.325	0.381	0.425	0.444	0.461	0.473	0.477	0.395
4	0.250	0.303	0.341	0.365	0.382	0.391	0.396	0.307
2	0.186	0.231	0.273	0.284	0.300	0.298	0.305	0.231
1	0.138	0.174	0.195	0.221	0.240	0.230	0.240	0.170

Table 2.6: KL-divergence values for natural scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25
Image-specific						0.424
63	0.900	0.763	0.707	0.684	0.670,	0.662
32						0.678
16						0.707
8						0.757
4						0.850
2						1.037
1						1.467

Table 2.7: KL-divergence values for urban scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25
Image-specific						0.364
63	1.274	1.190	1.153	1.141	1.137	1.139
32						1.153
16						1.201
8						1.298
4						1.501
2						1.981
1						3.280

Chapter 3

Saccadic Momentum and Facilitation of Return Saccades Contribute to an Optimal Foraging Strategy.

This article has been published in the journal PLoS Computational Biology: Wilming, N., Harst, S., Schmidt, N., & König, P. (2013, January). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Computational Biology*, *9*(1), e1002871

3.1 Abstract

The interest in saccadic IOR is funneled by the hypothesis that it serves a clear functional purpose in the selection of fixation points: the facilitation of foraging. In this study, we arrive at a different interpretation of saccadic IOR. First, we find that return saccades are performed much more often than expected from the statistical properties of saccades and saccade pairs. Second, we find that fixation durations before a saccade are modulated by the relative angle of the saccade, but return saccades show no sign of an additional temporal inhibition. Thus, we do not find temporal saccadic inhibition of return. Interestingly, we find that return locations are more salient, according to empirically measured saliency (locations that are fixated by many observers) as well as stimulus dependent saliency (defined by image features), than regular fixation locations. These results and the finding that return saccades increase the match of individual trajectories with a grand total priority map evidences the return saccades being part of a fixation selection strategy that trades off exploration and exploitation.

3.2 Author summary

Sometimes humans look at the same location twice. To appreciate the importance of this inconspicuous statement you have to consider that we move our eyes several billion (10^9) times during our lives and that looking at something is a necessary condition to enable conscious visual awareness. Thus, understanding why and how we move our eyes provides a window into our mental life. Here we investigate one heavily discussed aspect of human's fixation selection strategy: whether it inhibits returning to previously fixated locations. We analyze a large data set (more than 550,000 fixations from 235 subjects) and find that, returning to previously fixated locations happens much more often than expected from the statistical properties of eye-movement trajectories. Furthermore, those locations that we

return to are not ordinary – they are more salient than locations that we do not return to. Thus, the inconspicuous statement that we look at the same locations twice, reveals an important aspect of our strategy to select fixation points: That we trade off exploring our environment against making sure that we have fully comprehended the relevant parts of our environment.

3.3 Introduction

The effect of inhibition of return (IOR) was first described by Posner and Cohen (1984). When (covert) attention is attracted by a peripheral cue, reaction times to a subsequent probe stimulus in the same location depend in an intriguing way on the temporal offset between cue and probe: When the probe follows the cue at temporal offsets shorter than 225ms, fast responses are observed. In contrast, longer offsets (225-1500ms) lead to prolonged response times. In the original experiment, a central cross-had to be fixated continuously, so the inhibitory influence at long stimulus intervals pertained to covert attention. Along similar lines, overt attention—i.e., eye movements—shows the effect of temporal IOR as well. Specifically, the fixation duration before a return saccade is on average longer compared to a saccade that continues in the same direction as the previous one. Unfortunately, several conflicting results make a comprehensive explanation of saccadic IOR and its function difficult. This study aims at a step towards an understanding of these conflicting results by further characterizing the properties of return saccades and by providing a novel view of IOR during viewing of pictures of natural and urban scenes.

But first, we shortly recap some of the discussion surrounding a functional interpretation of IOR. Posner and Cohen hypothesized that IOR might prevent the return of attention to already processed locations. A further investigation by Klein and MacInnes (1999) revealed that eye movements are spatially biased away from the last (1-back) and second to last (2-back) fixation locations. This

established the interpretation of saccadic IOR not only in the form of a delay, but also in spatial terms as a "foraging facilitator". That is, the function of saccadic IOR is to direct attention to unexplored parts of the stimulus, thereby fostering optimal foraging behavior. This conjecture subsequently found its way into computational models of fixation selection where saccadic IOR prevents fixating on a location twice (Itti & Koch, 2001b; Parkhurst et al., 2002; Peters et al., 2005; Zhang et al., 2008; Zelinsky, 2008).

Whether saccadic IOR supports such a functional "facilitator" role has been heavily discussed. There is conflicting evidence on the spatial properties of return saccades. Several studies (Klein & MacInnes, 1999; Hooge et al., 2005; Smith & Henderson, 2009, 2011b, 2011a) have investigated how often return saccades occur and found, depending on the precise comparison, an elevated or attenuated number of return saccades. Thus, although of crucial importance for the functional interpretation of saccadic IOR, its spatial properties are still hotly debated. There is also mixed evidence on the temporal properties of IOR. Several studies report a significantly prolonged duration of fixation before saccades to the last fixation location (Klein & MacInnes, 1999; Hooge et al., 2005; Hooge & Frens, 2000). However, (Smith & Henderson, 2009; Anderson, Yadav, & Carpenter, 2008) reported a general dependency of fixation durations on the angular difference between the previous and the next saccade (termed "saccadic momentum" by Smith and Henderson). They argue that this accounts for parts of temporal IOR but that an additional localized inhibition zone remains. For saccades to the penultimate (2-back) fixation location conflicting evidence is reported whether 2-back return saccades are delayed (Klein & MacInnes, 1999; Hooge et al., 2005; Smith & Henderson, 2009, 2011a). In summary, the conflicting evidence of temporal and spatial properties makes it difficult to interpret saccadic IOR as a "foraging facilitator".

The dominating suggestion in the literature is that IOR supports optimal foraging strategies. This is fueled by the intuition that returning to previously fixated locations is not optimal for foraging

because a return saccade does not explore new parts of the environment. Hence, alternating observations of the presence/absence of inhibition of return have been taken as evidence in favor/against an optimal search strategy. However, these arguments are typically based on implicit assumptions regarding an optimal strategy and laboratory experiments with a task where it is difficult to identify the optimal foraging strategy, and therefore not based on direct investigations of fixation selection strategies. Therefore, it is presently unclear whether return fixations, contrary to the assumption that they are non-optimal, can actually be part of an optimal fixation selection strategy under natural conditions. With this in mind, we arrive at the key question of whether return locations are different from other fixation locations. For example, especially salient locations might be more likely to be fixated again, or targets of return saccades might require significantly more time to be comprehended compared to normal fixations. Such findings would suggest that return saccades might actually be due to a fixation selection strategy that needs to find a trade-off between factors such as exploration and comprehension.

We present a thorough investigation of temporal and spatial properties of return saccades by evaluating a large eye-tracking data set compiled from a host of different studies (Açik et al., 2010; Kaspar & König, 2011a, 2011b; Wilming, Betz, Kietzmann, & König, 2011). We analyze more than half a million fixations collected with natural scenes, urban scenes, fractals and pink noise images from 235 subjects in 5 different studies. These studies employed either free viewing conditions or a delayed patch recognition task. First, we analyze the frequency of 1- and 2-back return saccades and compare them to estimates of the number of return saccades expected from the statistical properties of single saccades and saccade-pairs. We also investigate the temporal properties of return saccades—i.e., if they are preceded by prolonged fixation durations—while paying attention to the effect of saccadic momentum. We then investigate the relationship of return locations to bottom-up saliency (as defined

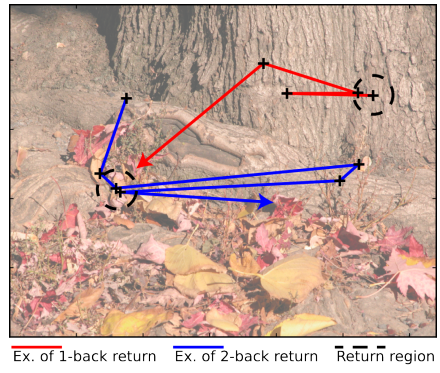


Figure 3.1: Figure 1. Example image from the category 'natural scenes'. The red line represents part of a trajectory that contains a 1-back return saccade. The blue trajectory contains a 2-back return saccade. The return region, used as a definition for return saccades for the temporal, saliency, and fixation sampling analysis is marked by the dashed circle.

by local image properties). Finally, we investigate the functional role of return saccades to get a better understanding of the functional purpose of saccadic IOR and what exploration strategies could lead to the observed pattern of return saccades.

We arrive at the view that saccadic momentum can fully account for temporal IOR; that return locations are highly salient and warrant increased scrutiny by the human observer; that this scrutiny is implemented by return saccades that are observed more often than expected by chance and by increased fixation durations at return locations; and that these properties of return saccades contribute to an optimal explorative strategy.

3.4 Results

Spatial Properties of Return Saccades

We started by investigating how often return saccades occur during viewing of natural scenes. Figure 1 shows an example image with

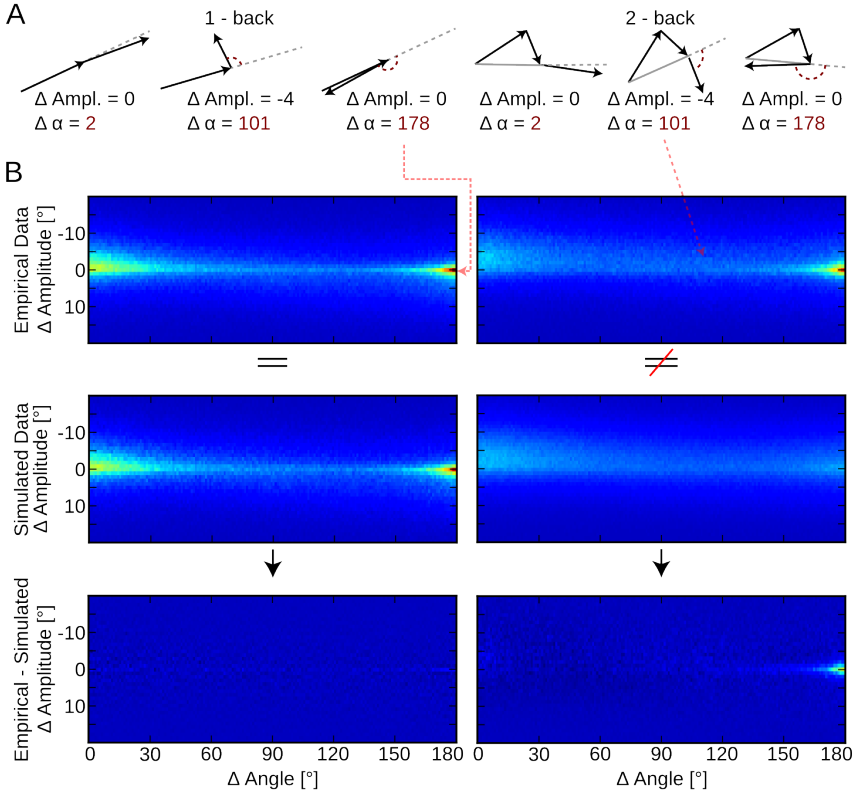


Figure 3.2: Figure 2. A shows an iconic depiction of forward (ΔAngle $\sim 0°$), perpendicular (ΔAngle $\sim 90°$) and return saccades (ΔAngle $\sim 180° \wedge$ ΔAmplitude $\sim 0°$) in the 1 and 2-back case and their associated angle and amplitude differences. The first row in B shows the distribution of amplitude and angle differences for empirical 1-back (left) and 2-back (right) saccades. In both cases, a pronounced return saccade peak is observable. The second row shows the same, but for saccades generated with our saccade simulator. Notably, the return peak for 1-back saccades matches the peak in the empirical data while the 2-back return peak is not reproduced. The difference between empirical data and simulator output is shown in the third row (same color scheme as above). The comparison of 2-back saccades shows systematic deviations for return saccades.

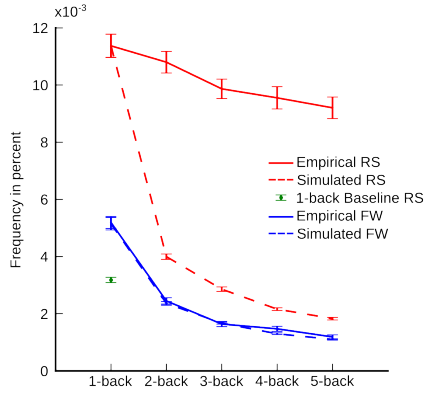


Figure 3.3: Figure 3. The frequency of return and forward saccades in empirical data and in simulation. For the case of 1-back saccades, the number of empirically observed return saccades ('Empirical RS') is larger than expected by chance ('1-back Baseline RS') and larger compared to the number of forward saccades ('Empirical FW'). The simulator reproduces the number of forward ('Simulated FW') and return saccades ('Simulated RS'). In the case of 2-back saccades, we find more return saccades than expected from the statistics of 1-back saccades, while the number of forward saccades is identical to the number of simulated forward saccades. When analyzing the presence of 3- to 5-back saccades, a similar pattern holds. Errorbars are bootstrapped 95% confidence intervals.

a 1-back (red) and a 2-back (blue) trajectory. Figure 2B (top left row) shows the frequency of saccade pairs with a specific amplitude and angle difference. In this plot, return saccades have a value of $\Delta\text{Angle} = 180°$ and $\Delta\text{Amplitude} = 0°$. We compared the number of 1-back return saccades to either the number of forward saccades or to a shuffled baseline (Hooge et al., 2005) that preserved the distribution of saccade amplitudes and angles but removed order effects. The shuffled baseline accounts for return saccades due to preferences of saccade angle and amplitude combinations by the oculomotor system, but does not contain return saccades caused by facilitation or inhibition of return.

In both cases, we found significantly more return saccades (bootstrapped, $p < 0.001$, Figure 3, 95% CIs created by bootstrapping per-subject percentages) in the empirical data than in the 1-back baseline. Qualitative inspection of the distribution of angle- and amplitude-differences (see Figure 3B top left row) revealed a sharp peak for return saccades ($178° \leq \Delta\text{Angle} \leq 180°$,$-1.5° \leq \Delta\text{Amplitude} \leq 1.5°$) while forward saccades ($0° \leq \Delta\text{Angle} \leq 2°$,$-1.5° \leq \Delta\text{Amplitude} \leq 1.5°$) appeared frequently but covered a larger range of amplitude and angle differences. We also observed an asymmetry with respect to amplitude-differences. Forward saccades were often shorter than their preceding saccades (see Figure 2B top left panel). In summary, 1-back return saccades appeared much more often than expected by the distribution of saccade amplitudes and angles, and even more often than forward saccades.

Next we investigated how often 2-back return saccades occur during viewing of natural scenes. While shuffling the order of saccades removes order effects for 1-back return saccades, it does not produce an adequate control distribution for 2-back return saccades. In order for this to be the case, one has to keep all 1-back return saccades due to preferences of the oculomotor system for combinations of angle and amplitudes of two consecutive saccades, but ignore all effects due to preferences of the oculomotor system for angle and amplitudes between three or more consecutive saccades (see Figure 2A, 2-back). We created control trajectories by sampling of saccades from the conditional distribution $P(L_{t+1}, \Delta\alpha_{t+1}|L_t)$ (see Materials and Methods) for each subject. This distribution expresses the probability of a saccade with amplitude L_{t+1} and angle difference $\Delta\alpha_{t+1}$ given that the last saccade had amplitude L_t. It fully characterizes the angle and amplitude dependencies between two consecutive saccades but does not contain information about 2-back return saccades. To create a trajectory, we randomly drew a saccade from the distribution of first saccades for a given subject and then determined the next saccade's angle and amplitude by sampling from $P(L_{t+1}, \Delta\alpha_{t+1}|L_t)$. We then iteratively added saccades to the

trajectory by sampling new amplitudes and angle differences from $P(L_{t+1}, \Delta\alpha_{t+1}|L_t)$, always reusing the last angle and amplitude. We matched the length of the simulated trajectories to the empirically observed lengths'.

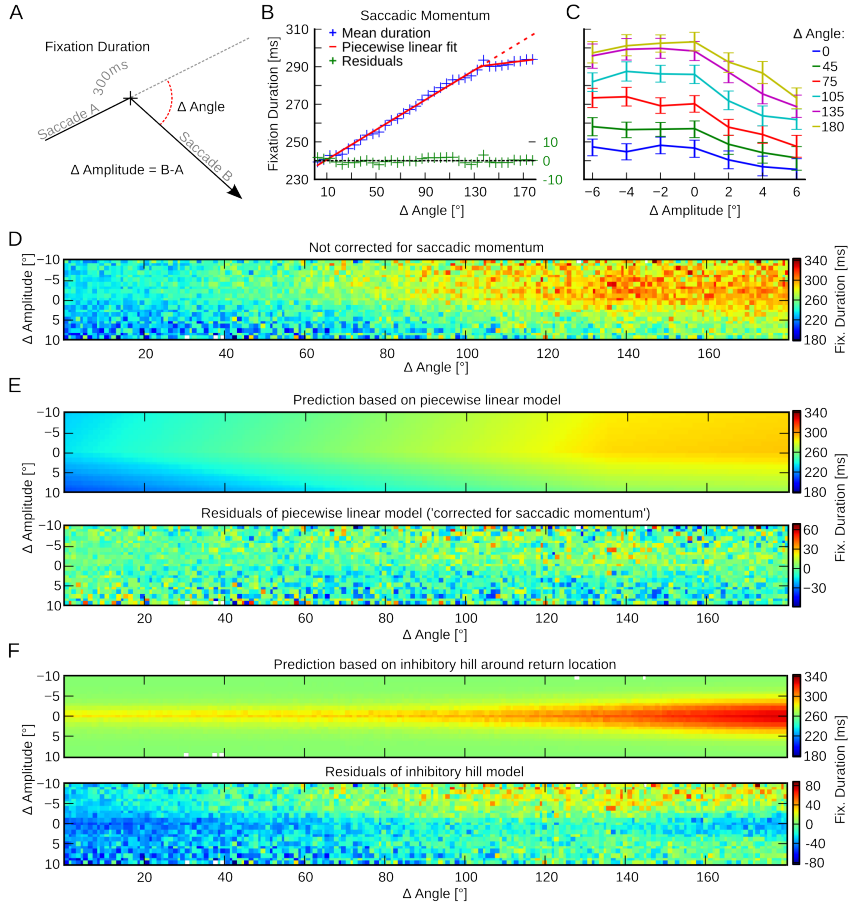
The control trajectories reliably reproduced 1-back dependencies and the number of 1-back return saccades in particular, as well as the overall shape of the distribution of angle- and amplitude-differences between consecutive saccades (see Figure 2B, left panels). However, the control trajectories contained fewer 2-back return saccades than observed in the real data (0.0040, bootstrapped CI [0.0038, 0.0042] vs.0.0108, bootstrapped CI [0.0100, 0.0116], Figure 3). The number of 2-back return saccades was much larger than the number of forward saccades (0.0023, CI [0.00211, 0.00244], Figure 2). In fact, the 2-back histograms of the simulated and the empirical data were very similar, with the exception of the return saccade peak. We thus conclude that the statistical structure of three consecutive saccades can be explained entirely from the statistical structure between pairs of saccades, with the exception of the increased amount of return saccades.

Despite the fact that the control trajectories do not preserve statistical effects of saccade triplets and saccade quadruples, we still compared the number of 3- and 4-back return saccades to the number computed from the control trajectories. In all cases, we observed many more return saccades in the empirical data (see Figure 3). We also found more return saccades than empirical forward saccades for 3- and 4-back saccades.

We conclude that locations that have been visited before are likely to be re-fixated, and for longer trajectories, this cannot be explained by the conditional dependencies between two consecutive saccades alone. We find that 1- to 4-back return saccades occur much more often than expected, but we do not observe any deviations from the predictions based on the statistics of saccade pairs for other saccades.

Temporal Properties of Return Saccades

After investigating spatial properties of return saccades we turned to temporal properties. The investigation of temporal IOR is complicated by a dependence of fixation duration on the angle and amplitude difference between the incoming and the outgoing saccade (see Figure 4B and also Smith and Henderson, 2009, 2011b, 2011a). On average, it takes longer to initiate a saccade perpendicular to the last saccade relative to a forward saccade, an effect termed 'saccadic momentum'. Because this effect is reminiscent of classical IOR effects we wanted to explicitly account for saccadic momentum. To achieve this we fitted a piece-wise linear model to the fixation duration data of each subject. In a fixation sequence A->B->C the model predicted the fixation duration at location B based on the amplitude and angle differences between saccades from A->B and B->C. We used a piecewise linear model with two slopes for angle and amplitude differences respectively (Figure 4B,E). The slopes for angle differences changed at a critical angle that was fitted at the same time. However, the position of the slope change for amplitude differences was set to 0°. Please note that for visualization purposes Figure 4 shows models fitted on all data, but for the analysis models were fitted for each subject individually with a least squares procedure. The subject specific models accounted for 10% of the variance in the fixation duration data. Contrary to Smith and Henderson (2009) we found that saccadic momentum did not increase linearly with the angle difference, but exhibited a change in slope for angle differences larger than 117° (CI [109, 124], slope of first segment 0.383 ms/°, CI [0.350, 0.416], and slope of second segment 0.002 ms/° CI [-0.13, 0.116], Figure 4A,B,E). The slope of the second segment is not significantly different from 0° and therefore indicates that no additional delay after the breakpoint at an angle difference of 117° occurs and that return saccades are faster than predicted by the first slope (see Figure 4B, compare red solid vs. dashed line). Thus, saccadic momentum is captured by a model with two different parts:


A
Fixation Duration
300ms
Saccade A
Δ Angle
Saccade B
Δ Amplitude = B-A
B
Saccadic Momentum
Mean duration
Piecewise linear fit
Residuals
Fixation Duration [ms]
Δ Angle [°]
C
Δ Angle:
0
45
75
105
135
180
Δ Amplitude [°]
D
Not corrected for saccadic momentum
Δ Amplitude [°]
Δ Angle [°]
Fix. Duration [ms]
E
Prediction based on piecewise linear model
Residuals of piecewise linear model ('corrected for saccadic momentum')
F
Prediction based on inhibitory hill around return location
Residuals of inhibitory hill model

up to angle differences of 117° fixation duration increases with 0.383 ms/° but larger angle differences do not incur a larger delay. We hypothesize that the different slopes might be due to two mechanisms that contribute to eliciting saccades with different dependencies on relative angle.

Amplitude differences changed slope at 0°, undershooting saccades had a slope of 0.39 ms/° CI [0.18, 0.60] and overshooting saccades had a slope of -2.75 ms/° CI [-3.02, -2.50].

In conclusion, the shallow slope for undershooting saccades, together with the position of the angle difference breakpoint at 117° and the 0 ms/° slope afterwards, show that the saccadic momentum effect is not specific to the return location.

Additionally we investigated IOR, similar to Smith and Henderson (2009) by comparing over- and under-shooting saccades with an

Figure 3.4 *(preceding page)*: Figure 4. The saccadic momentum effect. A) Schematic drawing of plotted fixation durations, angle, and amplitude differences. B) Average fixation durations, corrected for the effect of saccade amplitude difference, as a function of the angle difference between two saccades (data is pooled over all subjects). Turning the direction of a saccade prolongs the fixation duration before the saccade is made. C) Shows average fixation durations for specific combinations of amplitude and angle differences (data is binned with bin sizes of 30° and 2° for angles and amplitudes respectively; errorbars are 95% CIs over subjects). This shows that there is no increase of fixation duration for return saccades, except for the effects of angle and amplitude differences. D) Same as C but with bin sizes of 1°; fixation durations are color-coded. E) Top panel: Prediction of average fixation duration based on the piecewise linear model (the fit is based on pooled data over all subjects for visualization purposes). Bottom panel: Residuals of correcting for angle and amplitude differences with the piecewise linear model. Here the fit was done for each subject individually, and we averaged after the correction. F) Top panel: Prediction of average fixation duration based on the inhibitory hill model (the fit is based on pooled data over all subjects for visualization purposes). Bottom panel: Residuals of the inhibitory hill model. Here the fit was done for each subject individually, and we averaged after the correction.

angle difference of 180±30°. Contrary to Smith and Henderson (2009) we found no sign of a prolonging of exact return saccades (see Figure 4C, CIs for ΔAngle $\sim 180°$, $-6° \leq \Delta$Amplitude $< 2°$ largely overlap) compared to undershooting saccades. To exclude a potential effect of binning, we repeated this analysis with bins that were only one degree wide (see Figure 4D).

We also wanted to rule out the possibility that, additional to the spatially unspecific saccadic momentum effect, a spatially specific temporal IOR effect existed. We therefore fitted an 'inhibitory hill' model to the data (Figure 4F). Similar to the piecewise linear model, in a triplet of fixations A->B->C, we predicted the duration of fixation B. But this time we assumed that a Gaussian like inhibitory hill centered on fixation A would increase the duration of fixation B. The size of the inhibition was proportional to the distance between fixations A and C. We fitted this model with a least squares procedure with inhibitory hills of different sizes. The best fitting model had a Gaussian inhibitory hill with $\sigma = 3.12°$ and explained 1.6% of the variance in the fixation durations. When we fitted the same model on the residuals of the piecewise linear model the variance explained dropped to less than 0.0001%. Hence, on its own the 'inhibitory hill' explains much less variance of the data than the piecewise linear fit and adding the 'inhibitory hill' model to the piecewise linear fit had virtually no benefit. We conclude that the residuals of the piecewise linear model do not contain an effect of temporal inhibition of return anymore.

Next, we investigated the effects of correcting for saccadic momentum with our piecewise linear model. Figure 4E (bottom panel and Material and Methods) shows the residuals of the piecewise linear model. We observe that fixation durations before return saccades are not systematically different from fixation durations before saccades to other locations (see Figure 4E bottom panel). In contrast, the residuals of the inhibitory hill model (Figure 4F bottom panel) show systematic dependencies on angle and amplitude differences between saccades.

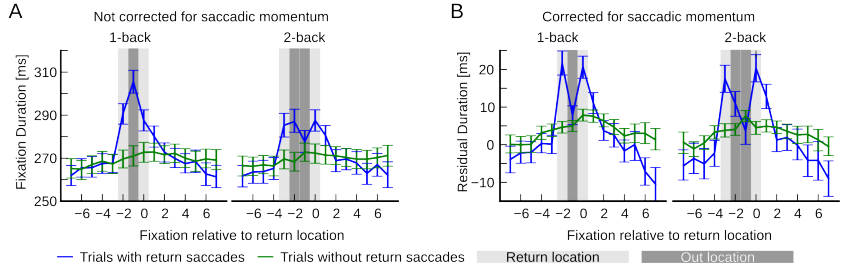


Figure 3.5: Fixation durations at return locations are longer. A) The average fixation duration at return locations in trials with 1-back and 2-back return saccades (blue lines) is longer than fixation durations in control trajectories (green lines). Errorbars are 95% CIs bootstrapped over subjects. B) Correcting for saccadic momentum with the piecewise linear model completely removes any trace of temporal inhibition of return for 1- and 2-back return saccades. Errorbars are 95% CIs bootstrapped over subjects.

In summary, the prolonging of fixation durations before return saccades can be explained in terms of saccadic momentum and saccadic momentum is not specific to return locations.

We next considered fixation durations at return locations to investigate if they are looked at more often because they were not scrutinized sufficiently the first time around (Hooge et al., 2005) or because they are highly salient and also demand above-average processing time.

To this end, we compared all trials (i.e. the entire fixation trajectory of one subject on one image) that contained return saccades (RS-trials) with all trials that contained no return saccade. We centered all RS-trials on the 2nd fixation of the return location. We aligned trials of the same length without RS to the trials that contained a RS. If for example, the 2nd fixation of the return location occurred at fixation Nr. 5, both trials were centered on fixation Nr. 5. Figure 5A shows that fixation durations at the return location are significantly longer than at control locations. Remarkably, this even holds when the location is visited for the first time. We observed the same

pattern with a reduced effect size for 2-back return saccades. Hence, return saccades are not due to a shortened analysis at first fixation; fixation duration is significantly increased during first fixation and re-fixation.

Please note that correcting for saccadic momentum and saccade amplitude differences eliminates the increase in average fixation duration before the return movement, where IOR has been typically observed (see Figure 5B). This supports our conclusion that controlling for the effects of saccadic momentum explains the prolonging of fixation durations before return saccades in our data.

To check if saccadic IOR effects that could not be explained by saccadic momentum were present in the individual experiments that we analyzed, we repeated the comparison of RS-trials and non-RS trials for every dataset. We checked if the difference at the out-location between both trial types was significantly different from zero when we corrected for saccadic momentum with our piecewise-linear model. We did not find any significant deviations (paired T-test, $p > 0.05$, Bonferroni corrected).

In summary, in our data temporal effects of IOR can be accounted for by a pronounced, non-linear effect of saccadic momentum and saccade amplitude differences, which is not specific to the return location. Additionally, the average fixation duration at the return location is longer for 1- and 2-back saccades, already during the first visit.

Return Saccades and Saliency

The observation of increased fixation duration at return locations suggests that such locations are special. To investigate whether the stimulus was systematically different at return locations compared to regular fixations, we computed bottom-up saliency at both locations based on the values of a large number of low (e.g. luminance, red-green and blue-yellow contrast) and mid-level (e.g. symmetry, intrinsic dimensionality) stimulus features.

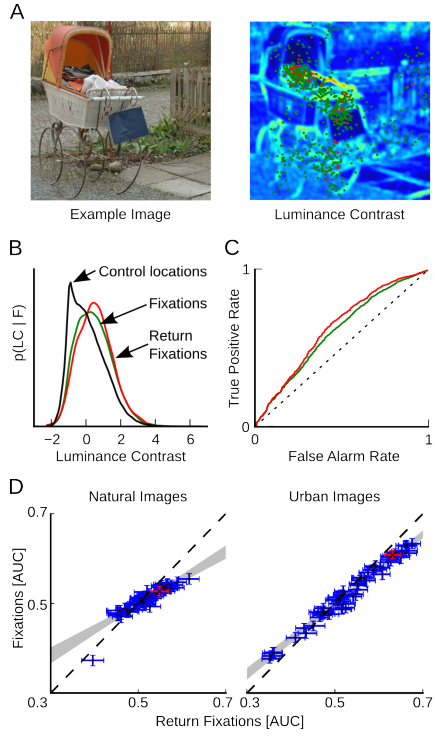


Figure 3.6: Image features predict return locations better than normal fixation locations. A) The right panel shows the luminance contrast feature for the image on the left. Green dots mark regular fixations, and red dots mark return locations. B) The distribution of feature values at control locations, regular fixation, and return fixation locations. C) The ROC curve for separating regular and normal fixations from control locations. D) AUC values of individual image features for return and regular fixation locations. Return locations are systematically better predicted by image features than regular fixations—i.e., return location feature AUCs are higher for predictive features (AUC $>$ .5) and smaller for anti-predictive features (AUC $<$ .5). Error bars are bootstrapped 95% CIs. The relationship between regular feature-fixation AUCs and return feature-fixation AUCs is well described by a linear relationship $\langle r^2 \rangle = 0.85$, $\langle \beta \rangle = 0.66$. Gray shaded area: convex hull of regression fits between return and regular feature AUC patterns.

We compared the values of 63 local features (please see Materials and Methods: Feature Analysis for the complete list) at return and non-return (normal) fixations in the dataset used in (Wilming, Betz, Kietzmann, & König, 2011) (Figure 6A,B). For quantification, we computed the area under the receiver-operating characteristics curve (AUC) of a linear classifier that separates return and normal fixation locations from control locations on the same image (Figure 6C) (Tatler et al., 2005). The AUC measures how well a feature can be used for correct classification. 0.0 implies perfect classification but switched labels; 0.5 is chance performance; 1.0 is perfect. Control locations were sampled from all fixation locations made on other images from the one in question and hence take into account the general spatial bias. We calculated the AUC for separating return fixation locations from controls and the AUC for separating normal fixation locations from controls.

We observe a linear relationship between AUCs of different features calculated for return-locations and normal-fixations. Furthermore, this holds for natural and urban scenes (Figure 6D, each data point shows AUC values for one image feature). The pattern of AUC values for return and normal fixations is well described by a linear relationship (natural scenes: $r^2 = 0.76$, urban scenes: $r^2 = 0.95$). Only the phase congruency feature does not fit this linear pattern; it is slightly better for predicting normal fixations than return fixations (lower left corner in left panel of Figure 6 D). Importantly, the slope of the linear fit is less than 1.0 (natural scenes: $\beta = 0.56$, T-test $\beta < 1 : p < 0.0001$; urban scenes: $\beta = 0.77$, T-test $\beta < 1 : p < 0.0001$). Hence, those features that predict normal fixation locations above chance (AUC>.5) better predict return locations than regular fixation locations. Importantly, those features that are anti-predictive (AUC<0.5) are also more anti-predictive of return locations than of regular fixation locations. This indicates that the pattern of contribution of different features, as quantified by the AUC values, does not differ between normal and return locations. Such a linear relationship implies that image feature based salience

models trained only on regular fixation locations will perform better on return locations than on regular locations. In summary, image features better predict return locations than regular locations.

To compare bottom-up saliency values at return and normal fixation locations, we compiled a weighted sum of all 63 features into a single saliency score. Weights for the linear combination were obtained by a logistic regression that separated either return locations from controls (RS-model) or normal fixations from controls (FIX-model, see Materials and Methods). We then computed the AUC of both saliency scores for separating return-locations and non-return locations from controls. We used leave-one-subject-out cross validation to ensure independence of training and test data. We found that return locations could be better predicted (average AUC of 0.733; RS-model: 0.731, FIX-model: 0.736) compared to normal fixations (average AUC of 0.670; RS-model: 0.667, FIX-model: 0.674). From this analysis, we conclude that return saccades are directed to more salient locations than normal saccades and that the pattern of feature-fixation correlations is comparable to return locations and normal fixations.

Fixation Sampling Strategies

The finding of an increased number of return saccades and prolonged fixation durations at return locations is difficult to reconcile with a foraging strategy that maximizes the entropy of a fixation density map, i.e. the area that is 'covered' by fixations. Yet return locations are 'special' in the sense that they are looked at longer and do not appear at random locations. Instead of maximizing entropy, we hypothesize that the very existence of return fixations serves to optimize the match of saccadic trajectories with an internal priority map that encodes which locations are relevant in the scene. Here we replace the spatially flat prior of the maximal entropy assumption (Figure 7A) with a stimulus-dependent prior and use the viewing behavior of (other) subjects as a proxy for such an internal priority

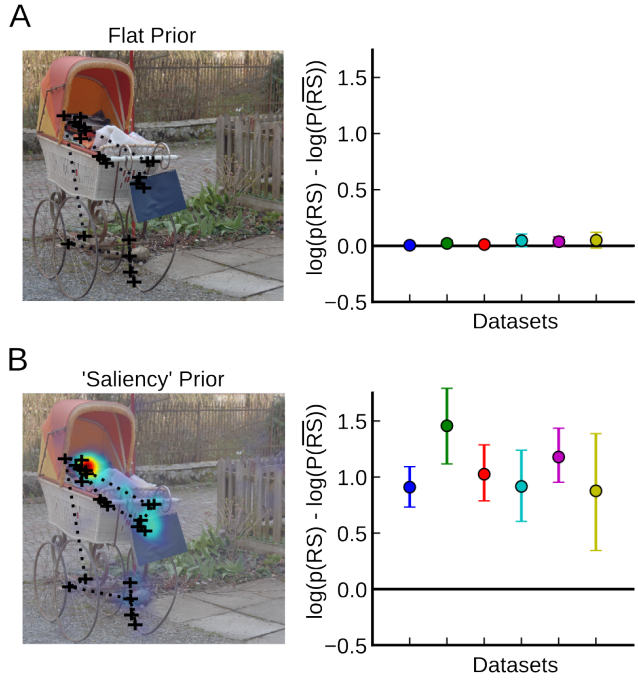


Figure 3.7: Return saccades increase the trajectory likelihood when observers sample from an empirical saliency distribution. Both plots show the difference of the log likelihood for trajectories with and without return saccades as a function of the dataset. A) If all locations have equal probability of fixation, trajectories with return saccades are as probable as trajectories without return saccades. B) If salient locations are more probable than other locations, trajectories with return saccades are more likely than others. Error bars are bootstrapped 95% CIs.

map (Figure 7B). That is, we use empirically defined salience, given by how often different subjects look at a location, as the internal priority map.

In this respect, we were interested if, all else being equal, a return saccade would increase the probability of a trajectory according to the internal priority map. We compared trajectories with return saccades to the same trajectories that, instead of exploiting an already seen location, explored one additional new location.

More specifically, for each fixation trajectory that contained a return saccade, we first computed a fixation density map from the fixations of all other subjects on the same image. We made sure that in this computation, trials containing return saccades were omitted (see Materials and Methods). We then used this fixation density map as an internal priority map for the trial in question. We compared the probabilities of generating two different kinds of trajectories based on the fixation trajectory in question from this internal priority map: The first contained the return saccade (return-trajectory) but we removed the last fixation. For the second trajectory (exploration-trajectory) we removed the 2nd visit to the return location but kept the last fixation. The exploration and return trajectories thus contained the same number of fixations, but the exploration trajectory contained one more unique fixation location (see also Materials and Methods). In other words, given the original fixation trajectory A-B-A-C … -F-G, the return trajectory is given by A-B-A-C…F and the exploration-trajectory is A-B-C…-F-G.

The probability for the exploration and return-trajectories was defined as the probability to draw exactly these trajectories from a multinomial distribution with event probabilities given by the internal priority map. Because we use a multinomial distribution as our model, the order of fixations is irrelevant and changed distances between fixations do not confound the results. We find that return saccades actually increase the probability of a trajectory compared to the omission of such saccades (Figure 7, ANOVA with factors experiment and saliency map type, main effect of saliency map type

$p < 0.0001$, no other significant effects at $p = 0.05$).

In summary, to match an internal priority map it is better to allow return saccades to exploit empirically salient locations in the priority map compared to forcing all saccades to unexplored locations. This result is also reflected in the additional finding that return locations show higher average values of the internal priority map compared to locations before and after return locations. That is, humans try to visit empirically highly salient regions but trade off exploitation and exploration by revisiting important parts of the stimulus.

Discussion

In this study, we investigated the spatial, temporal and functional properties of saccadic inhibition of return.

With respect to spatial properties, we find more 1-back and 2-back return saccades than expected from the distribution of saccade angles and amplitudes and relative angles and amplitudes. Also, our novel statistical model for 2-back return saccades reproduces the distribution of angle and amplitude differences of saccade triplets very well except for 2-back return saccades. This indicates that our model is adequate to explain higher order biases in saccade trajectories but that 2-back return saccades are facilitated compared to these higher order biases.

This agrees with findings from Hooge et al. (2005) who used a comparable baseline for 1-back return saccades. Smith and Henderson (2009, 2011b, 2011a) used two different baselines but find similar results. Compared to distance matched controls (e.g. saccades with $\Delta\text{Amplitude} = 0^\circ$ and $\Delta\text{Angle} = 90^\circ$) they report an equal or larger number of 1-back and 2-back return saccades. Compared to a baseline where the order of fixations is shuffled they report more 1-back and 2-back return saccades in their empirical data.

In disagreement with our results Bays and Husain (2012) argue that 1-back return fixations occur less often than should be expected. The critical difference to our study is the baseline used for comparing

the number of return saccades. Bays & Husain argue that saccade trajectories are not only influenced by oculomotor biases but also by the spatial distribution of salient locations in an image. They generate control trajectories that take both biases into account by sampling from the conditional probability distribution $P(x_t = X|x_{t-1} = Y)$, which expresses the probability to fixate location X given that the current fixation is at location Y. Importantly, the resulting trajectories contain more return saccades than their empirical data. Because the process that generated these trajectories did not take into account past fixations but still created more return fixations, Bays and Husain conclude return locations are actively inhibited. What could explain the differences between our and Bays & Husain's findings? There are several differences regarding the acquisition and analysis of eye-tracking data in our and Bays and Husain's study. First, Bays and Husain presented images for 20s while in the present investigation the presentation time was 6s or shorter. Fixation trajectories over repeated presentations of the same stimulus are partly overlapping (Kaspar & König, 2011b) supporting the argument that prolonged stimulus presentations might have an effect on the frequency of return-saccades. Second, we seem to observe a much more localized return peak (compare our Figure 2B and their Figure 1B). This percolates to differences in the definition of return saccades between Bays & Husain's and our work and thereby to a different estimate of the number of empirical return saccades. Third, to accurately estimate conditional probability densities a large amount of data is required. In the present study we opted for a reanalysis of several previously conducted studies resulting in a very large database. This allowed us to remove trajectories containing return saccades from the estimate of fixation probability densities. Fourth, typical laboratory setups with limited size monitors enforce saccades with larger than 90° turning angle in order to maintain the gaze within the monitor boundaries, part of these are classified as return saccades. To resolve these issues and reach a final conclusion more research is warranted.

With respect to the temporal properties of return saccades, we

find that direct return saccades are preceded by longer fixations than forward or perpendicular saccades. We therefore replicate a classical effect of saccadic inhibition of return. However, this effect is explained by saccadic momentum (Smith & Henderson, 2009; Anderson et al., 2008). A piecewise-linear dependence of fixation durations on angle and amplitude differences between the preceding and next saccade, i.e. large turns in eye movement direction are preceded by longer fixation durations compared to small turns. We find that exact return saccades do not take longer than undershooting return saccades, or saccades with an angle difference larger than $\sim 117°$. Correcting for the effect of saccadic momentum removes the delay imposed on direct return saccades compared to non-return saccades. Crucially, we find that the dependence of fixation durations is constant after a critical angle. That is, return saccades are faster than expected from the slope of saccadic momentum before the critical angle. We furthermore tested directly if a localized 'inhibitory hill' around the return location could explain our data and found that it performed worse than the piecewise linear model. Additionally, the inhibitory hill model did not explain variance in our data that was not explained by the piecewise linear model. In conclusion, we did not find a spatially localized inhibitory effect for return saccades in addition to saccadic momentum. We therefore conclude that our data is better described by a spatial facilitation of return, a delay increasing linearly with angle of changing gaze direction (saccadic momentum) up to a critical angle and constant thereafter.

Our results are compatible with many findings in the literature. We replicate the classical saccadic IOR effect (Klein & MacInnes, 1999; Hooge et al., 2005) on a large data set and provide a detailed description of saccadic momentum (Smith & Henderson, 2009, 2011b, 2011a; Anderson et al., 2008). However, we could not replicate Smith & Henderson's (Smith & Henderson, 2009) finding of an extra delay, in addition to saccadic momentum, for return saccades. Also, our results contrast with findings of (Hooge & Frens, 2000) who find a localized temporal zone of inhibition of 4º by asking their subjects

to carry out a pre-determined sequence of saccades. The difference between Hooge et al.'s and our results might be explained by the two very different tasks and stimulus arrangements. It is well known that the oculomotor system in the brain includes many different areas that are tightly coupled (Lynch & Tian, 2006). Carrying out pre-programmed saccade sequences might recruit neural substrate that elicits temporal inhibition of return. Hooge et al. suggest that the superior colliculus might be the neural substrate that causes effects observed in pre-planned saccade sequences. In contrast, free viewing, where fixation locations are selected based on local salience, oculomotor bias's and other top-down factors, might activate different networks that lead to different temporal properties of fixations. One candidate would be LIP, which has been implicated in computing a priority map (Bisley, 2011), which combines bottom-up and top-down information to guide selection of fixation targets during visual search.

An alternative non-exclusive explanation, that would incorporate both contradicting results, might be that precise saccadic IOR can be tuned by the visual system. This is supported by a study from Farrell, Ludwig, Ellis, and Gilchrist (2010) that shows that the classical IOR effect is adaptive to environmental statistics. However, because they did not explicitly investigate saccadic momentum, it remains to be seen what is modulated: return location unspecific saccadic momentum, return location specific IOR or both.

Interestingly we find that return locations are more salient, according to empirically measured as well as stimulus dependent saliency, than regular fixations. Hooge et al. (2005) find that the first visit of a return location is shorter than the second visit. They suggest that return saccades occur because the visual system did not have enough time to analyze a fixation location during the first visit. In our data return locations are fixated longer compared to control fixation locations during both visits. The visual system therefore has more processing time available for return locations than for regular locations. These findings suggest that return locations need to be

scrutinized in more detail than regular fixation locations.

We also found that return saccades increase the match of individual trajectories with a grand total priority map. The priority map was defined by empirical salience, i.e. those locations that are consistently fixated by many subjects. Because trajectories that contained return saccades were more likely than trajectories that explored a new location with every fixation, we suggest that return saccades are the consequence of a fixation selection strategy that samples relevant parts of a scene. Furthermore, because the internal priority map was defined by empirical salience, which we interpret as a proxy for behavioral relevance, return locations were more relevant than other locations. We therefore suggest that the fixation selection strategy trades off exploration of unseen relevant locations and exploitation of already seen relevant locations with return saccades.

What are possible mechanisms that could explain our findings?

With respect to saccadic momentum, the question arises whether the observed regularities could be an effect of the physical properties of eye-movement control. Different patterns of muscle movements are necessary for return saccades and forward saccades. Forward saccades require flexed muscles to be flexed more, while stretched muscles must be stretched more. Return saccades require an inversion of these muscle states: flexed muscles must be stretched and stretched muscles must be flexed. This might contribute to the observed differences in fixation durations. However, when talking about muscle effects, two things should be kept in mind: First, the temporal difference between the length of fixations before return and forward saccades is in the order of 50ms (Figure 4B). Bahill, Clark, and Stark (1975) show that saccades of up to 20° can be completed in less than 60ms and thus it is safe to assume that the time needed for the acceleration of the eye during a saccade is much shorter than 50ms. Therefore, differences in activation patterns on the muscular

level cannot explain the systematic increase of fixation durations observed here.

Second, Farrell et al. (2010) found that the temporal difference between return and forward saccades is modulated by the likelihood of a return saccade, with the effect eventually vanishing when return saccades are very likely. While this does not rule out a contribution of muscle effects to saccadic momentum, the least it demonstrates is that saccadic momentum can be modulated by factors that are independent of physical motor control.

Ludwig, Farell, Ellis, Gilchrist, and Farrell (2009) proposed that 'Inhibition of Return' can be explained in a decision-making framework. In short, potential saccade targets accumulate evidence until an evidence threshold is reached. The first target that reaches the threshold is used as the next saccade target. Indeed, they find a difference in saccade latency for return and non-return saccades that correspond to differences of accumulation rate in their fitted models. However, they only differentiate return and non-return saccades. Thus, the phenomenon of saccadic momentum makes a large contribution to that comparison and might easily dominate. It would be interesting to see if the accumulation rate is parametrically modified by the angle difference between the new saccade target and the last saccade. However, even if a change in accumulation rate can explain saccadic momentum, this would make the high incidence of return saccades even more puzzling. Furthermore, the question remains why the accumulation rate changes with changes in eye-movement direction.

But we find an alternative suggestion more tenable: If we imagine that we shift the center of gaze from point A to B, then parts of the stimulus between A and B will have been sampled by the fixation of A. Thus, relative to B, backward targets are at locations for which prior information exists while forward targets deal with parts of the stimulus for which no (or less) information is available at that moment in time. We hypothesize that forward and backward targets have different accumulation rates because different amounts of

knowledge are available for these locations. Considering that receptive fields are remapped during saccades, it does not seem unlikely that prior knowledge is transferred during the remapping (Hall & Colby, 2011). For salient targets to reach the decision threshold faster when they are 'forward' compared to 'backwards', the accumulation rate has to be inversely proportional to the amount of knowledge available. This would imply that more prior knowledge leads to slower accumulation of evidence. It seems that such a notion is compatible with accounts of predictive coding in which higher-level information explains away activity at lower levels (Rao & Ballard, 1999). Here the higher-level knowledge about backwards locations 'explains away' their salient properties, thereby making them less salient compared to forward locations. This in turn would lead to a slower accumulation of evidence at backward compared to forward locations. In that context, the observed piecewise linear dependence of fixation duration on saccadic angle is important. Based on this observation we suggest that two competing mechanisms are active in parallel, and only one of these—the one with a steep dependence of fixation duration on the saccadic angle—is dependent on the already available knowledge. While this is speculative, the proposal does fit with our finding of increased bottom up saliency at return locations.

In summary, accumulator models are a promising tool to understand the dynamics of saccade target selection. Future studies will need to link saccadic momentum and facilitation of return to specific properties of such models. The findings that return locations are more salient and looked at longer must be crucial parts of this puzzle.

What could be the function of facilitation of return and delay of direction change?

Clearly, spatial facilitation of return is incompatible with the objective of covering the entire stimulus evenly with fixations in a short

amount of time. However, what is the motivation to assume that the stimulus is equally interesting in all locations? In an everyday search task, such as when looking for the car keys, one would not cover all places from cellar to rooftop evenly. Instead, it is sensible to scrutinize those locations that are likely due to prior knowledge and to look twice before considering more exotic alternatives. Under laboratory conditions, for example when a near threshold Gabor patch is superimposed on a pink noise image at a random location, the search strategy might adapt to the flat location prior (Najemnik & Geisler, 2005). This is a remarkable feat of behavioral adaptation, yet no reason to assume that return saccades are generally inhibited. During free viewing, no explicit external task is enforced and subjects do not relate their eye-movement behavior to an externally set optimality criterion. Some studies included in our data set employ specific tasks. Specifically, in the delayed patch recognition task (Açik et al., 2010) subjects have to decide whether a target patch was contained in a previously shown sample image. The target patches are selected uniformly from the entire stimulus, which might suggest that return saccades are not useful to solve the task. However, the probe patch is not presented in the location where it was in the stimulus and after stimulus offset only. This makes keeping track of where in the sample image the target patch was selected difficult. Furthermore, to prevent fatigue, Açik et al. (2010) deliberately choose to present only 128 images to each subject. The number of trials is therefore considerably smaller than in psychophysical studies with reduced setups. Hence, the opportunity for subjects to infer and adapt to the objective prior of patch locations is rather limited. But even for fully adapted subjects, it is unclear whether seeing the entire stimulus is the optimal strategy for a delayed patch recognition task. The task requires not only passive observation of the stimulus but encoding and recalling as much as possible of it at a later stage. The optimal strategy needs to trade off holding complex stimulus patches in memory with exploring new parts of the stimulus. In this respect, return saccades might be part of an optimal strategy because they

allow the visual system to exploit information at relevant locations more thoroughly.

It could be argued that we did not use a visual search task and therefore found more return saccades than expected. As described above two studies included in our data set employed a delayed template match search task where homogeneously distributed fixation locations seem advantageous. Furthermore, even during visual search return saccades are not automatically disadvantageous for search performance. First, a consistent central bias has been documented in many studies (for example Wilming, Betz, Kietzmann, and König, 2011; Tatler and Vincent, 2008), invalidating an assumption of a flat prior. This shows that the visual system does not consider every location to be equally relevant. Second, Hooge et al. (2005) find more 1-back return saccades than expected during a visual search task. Third, even during visual search, a single fixation might not suffice to identify a target in front of the background, and there is evidence that humans take uncertainty inherent in their visual system into account (Najemnik & Geisler, 2005). Also, there clearly are prior expectations about where targets of specific types can be found in a scene (Torralba et al., 2006) (e.g., pedestrians are usually not located in the sky). These two conditions necessitate trade off of exploration and exploitation in visual search—return saccades (exploitation) with saccades that target unseen parts of the stimulus (exploration). Therefore return saccades are likely to be a part of visual search strategies as well.

Having considered everyday search tasks, free viewing, and delayed patch recognition, we find it unconvincing that a flat spatial prior over stimulus locations must be part of a good strategy to solve these tasks. In turn, we argue that from the existence of return saccades, it does not follow that a task is not being solved optimally.

Instead, a novel view concerning the functional interpretation of IOR emerges. Farrell et al. (2010) have shown that the time difference between return and forward saccades is adaptive to the environment. Smith and Henderson (2011b, 2011a) argue that saccade latencies are

the result of several interacting processes such as bottom-up input, top-down control and saccadic momentum. We provide evidence for the hypothesis that return saccades are part of a strategy that aims at devoting attention to the most relevant information in the stimulus: First, return locations are more bottom-up salient than regular fixation locations, showing that the stimulus is different at return locations compared to regular locations. Second, return locations are fixated longer during both visits, indicating that more attention compared to regular fixation locations is devoted to return locations. Third, return saccades occur more often than expected, suggesting that they are an important part of a fixation selection strategy. Most importantly, if we accept eye movement behavior of other subjects as a proxy for relevance, return saccades increase the likelihood of a trajectory to sample the relevant parts of an image. We therefore conclude that spatial facilitation of return, saccadic momentum, and relative speed-up of saccades at very large angle differences might not serve a single objective but might emerge from the broader goal to optimally sample relevant parts of a stimulus.

3.5 Materials and Methods

Data

We re-analyzed data from several studies conducted at the Institute of Cognitive Science, University of Osnabrück. Here we briefly summarize the different studies but leave details to the respective original publications. Açik et al. (2010) investigated the effect of age on viewing behavior. They presented 128 images from the categories 'manmade scenes', 'natural scenes', 'fractals', and 'pink noise' for 5s. The images were selected from a larger database that contained 64 images per category. Images from the same database were used in (Kaspar & König, 2011a; Wilming, Betz, Kietzmann, & König, 2011). After each image, an image patch was presented, and subjects had to answer whether this patch was contained in the previously shown

image. Fifty-eight subjects participated in this study (18 elementary school children with a mean age 7.6, 23 university students with a mean age of 22.1, and 17 older adults with a mean age of 80.5). Wilming, Betz, Kietzmann, and König (2011) showed 128 images from the categories 'manmade scenes' and 'natural scenes' in a free viewing paradigm with a viewing duration of 6s to 48 subjects (aged 19 to 28 years). Kaspar and König (2011a) investigated the influence of repeated stimulus presentations, image category, and individual motivations. They presented 48 images taken from the same scene types used by Açik et al. (2010) and repeated the presentation of each image 5 times. The subjects were instructed to freely view the image for a period of 6s. Forty-five subjects participated in the study (aged 18–48 with a mean age of 24.2 years). Kaspar and König (2011b) (data from 'Experiment 2') presented 30 different urban scenes to 34 subjects (aged 19–49, mean age 25.9 years) with a viewing duration of 6s. Each image was presented five times to each subject. The images were not part of any of the other studies used here. We analyzed data from two more experiments; the results have so far not been published. In these studies, conducted by Alper Açik, 50 subjects were presented with contrast modified and phase scrambled images from the category 'fractals'. After a stimulus presentation of 5s, subjects had to perform a 2AFC patch recognition task (20 subjects) or a YES/NO patch recognition task (30 subjects). We treated the two different tasks as different datasets.

All studies used an Eyelink II eye-tracker (SR Research Ltd., Mississauga, Ontario, Canada). All studies were conducted in compliance with the Declaration of Helsinki as well as national and institutional guidelines for experiments with human subjects. Because different studies used different displays and image sizes, we converted all fixation coordinates into degrees of visual angle. In total we analyzed over 597,000 fixations collected from 235 subjects in 6 different studies.

Spatial Properties of Trajectories

To investigate the frequency of 1- and 2-back return saccades, we created two different baseline conditions. For the 1-back condition, we shuffled all of the recorded saccades. This removed all order effects but did not change the distribution of saccade angles and amplitudes. We used this shuffled baseline to estimate how many return saccades should be expected by randomly sampling from the distribution of saccade angles and amplitudes. All saccades with an angle difference larger than 178° and amplitude difference of less than ±2° were considered return saccades. To determine significant deviations of the number of return saccades from the shuffled baseline, we bootstrapped 95% confidence intervals around the mean difference of return saccades for each subject and checked if the confidence interval contained 0. In comparison, the empirical data contained significantly more return saccades in the 1-back condition. Bootstrapping the per-subject percentages created the 95% confidence intervals shown in Figure 3.

Subsequently, to investigate whether 2-back and higher dependencies between saccades can be explained by 1-back information, we devised a saccade generator, which uses 1-back information of trajectories as an input to generate arbitrarily long sequences of saccades. As the generator does not use any 2-back information, any patterns that can be observed in the 2-back condition of the generated data are due to 1-back dependencies alone. The generator creates a trajectory by drawing a saccade from the distribution of first saccades in the input data and copies its absolute angle and amplitude. Subsequently, further saccades are added by drawing their angle difference and amplitude with respect to the last saccade from the conditional distribution $P(L_{t+1}, \Delta\alpha_{t+1}|L_t)$. This distribution expresses the probability of observing an amplitude difference L_{t+1} and angle difference $\Delta\alpha_{t+1}$ at the next saccade, given that the length of the last saccades was L_t. It thus comprises only 1-back information. We estimated this distribution for every subject separately by

computing histograms for each possible value of L_t. Sampling from the non-conditional probability distribution from Figure 2 does not generate valid adjoining saccade trajectories because not all negative amplitude differences can be generated at all times. In terms of fixation coordinates, no additional restrictions were made such that the simulator precisely replicates 1-back dependencies without incorporating any additional image statistics such as picture size. The resulting set of fixations could be analyzed in terms of saccade dependencies equal to the empirical data. To validate the accuracy of the saccade simulator, we compared the similarity between subjects and the similarity between subjects and simulator. To this end, we computed for each subject the distribution $P_{emp}(\Delta L, \Delta\alpha)$ of amplitude differences ΔL and angle differences $\Delta\alpha$ (see Figure 2) for 1-back saccades. Subsequently we computed $P_{sim}(\Delta L, \Delta\alpha)$ for each subject based on saccades generated from their own distribution $P_{emp}(\Delta L, \Delta\alpha)$. Finally we computed the KL-divergence between subjects and between subjects and their simulated saccades:

$$\langle D_{LK}(P_{emp}(\Delta L, \Delta\alpha)^i || P_{emp}(\Delta L, \Delta\alpha)^j)\rangle \forall i, j \wedge i \neq j$$

$$\langle D_{LK}(P_{emp}(\Delta L, \Delta\alpha)^i || P_{sim}(\Delta L, \Delta\alpha)^j)\rangle \forall i$$

where i and j are subject indices. We found that the KL-divergence between subjects was higher than the divergence between subjects and simulator output. Additionally, the number of return saccades generated by the simulator is not different from the number of return saccades found in the empirical data. Furthermore, qualitative comparison of differences between empirical and simulated data did not reveal any systematic deviations in the 1-back case. From this, we conclude that the simulator reliably replicates all 1-back dependencies in the data.

To compare the number of return saccades, we again bootstrapped 95% confidence intervals around the difference of simulated and

empirical return saccades and checked if the interval contained 0. As expected, this was the case for 1-back saccades. All other comparisons showed significantly more empirical return saccades (see Figure 3). In the case of forward saccades, all comparisons contained 0.

To assess the similarity of the distributions $P_{emp}(\Delta L, \Delta\alpha)$ and $P_{sim}(\Delta L, \Delta\alpha)$ for the 2-back case (see Figure 2, left column), we calculated the KL-divergence between the two for each subject. The mean KL-divergence was 0.21, to which the return peak contributed more than any other area of comparable size (for example, 4 times as much as the forward peak). Thus, all other 2-back dependencies were very similar.

Temporal Properties of Return Saccades

Because effects of saccadic momentum on fixation durations are largest at return locations, they potentially confound findings of IOR. Smith and Henderson (2009, 2011a) considered the effect of saccadic momentum by comparing average fixation durations for exact return saccades and over- and under-shooting return saccades. We repeated this analysis but take several other measures to ensure a fair comparison. First, we explicitly estimated the effect of saccadic momentum and saccade amplitude differences on fixation duration with a non-linear breakpoint regression:

$$y = \beta_1\Delta\alpha + \beta_2(\Delta\alpha - s_{\Delta\alpha})k_{\Delta\alpha} + \beta_3\Delta L + \beta_4(\Delta L - s_{\Delta L})k_{\Delta L} + \beta_0$$

where $\Delta\alpha$ is the angle between the previous and next saccade, ΔL, is the amplitude difference, $\beta 1 - 4$ are the slopes of the individual linear segments, $s_{\Delta\alpha}$ is the critical angle, $s_{\Delta L} = 0$ and $k_x = \begin{cases} 1, & x < s_x \\ 0, & \text{else} \end{cases}$. The parameters β_{0-4} and $s_{\Delta\alpha}$ were fitted with a least squares procedure implemented in SciPy 0.9 for each subject individually. Please note that, for visualization purposes, the

model fit in Figure 4 was computed by using all of the available data. All other inferences are based on models that were fit on a per-subject basis. We chose a piecewise-linear regression for two reasons: First, the relationship of angle differences and fixation durations seem to exhibit two linear parts (see Figure 4B,D). That is, using a linear regression introduces systematically larger residuals for large angle differences and for small amplitude differences. This is potentially critical because according to our data, changes in slope are not specific to return locations, and thus do not represent a true IOR effect but instead might interact with inferences about effects of IOR. Second, the breakpoint regression is conceptually simple and provides a decent fit with the data ($r^2 = 0.1$, normalized RMSE = 0.16). Analyses that are 'corrected for the saccadic momentum effect' are carried out on the residuals of this regression.

Consecutively, we computed the duration of fixations with respect to the amplitude difference of the previous and consecutive saccade. Figure 4C shows an average over subjects for 30° bins for different saccade amplitude differences. Confidence intervals are based on bootstrapped across subject averages. Figure 4D shows the same but pooled over all subjects and for 1° bins for both angle and amplitude differences. Figure 4E shows the residuals of our piecewise-linear model, that is fixation durations corrected for saccadic momentum. Qualitative inspection shows that little structure remains in the residuals. Specifically, those areas where few samples are available (compare with Figure 2B, top left panel) show larger deviations than those where many samples are available. In an additional analysis (see Text S1) we found that such deviations can be expected even if no effect of angle and amplitude differences is present in the residuals.

Figure 5 shows trials with return saccades aligned to the return fixation and trials without return saccades. Trials without return locations were aligned as follows: For every subject we estimated $P(F_i|L)$ which expresses the probability that a fixation at location i within a trajectory is a return fixation given that the amplitude of the

trajectory is L. For every non-return trial we drew a return fixation location from this distribution and aligned the trial to this position. Error bars show bootstrapped 95% confidence intervals.

Feature Analysis

To assess the relationship of return locations and bottom-up saliency, we used a saliency model similar to Betz et al. (2010). We computed 63 different features that are predictors of fixation locations on plain RGB values of the images. We used luminance, saturation, blue/yellow color, and red/green color channels of the stimulus (Itti & Koch, 2001b). All features were computed on three different spatial scales, which were created by rescaling the input image with a Gaussian pyramid. For each feature on each spatial scale, we applied three different filters: Gaussian smoothing ($\sigma = 1.1°$), local contrast ($\sigma = 1.1°$), and texture contrast by calculating the local contrast twice on a feature ($\sigma_1 = 1.1°$, $\sigma_2 = 5.5°$). The local contrast is computed by $C = \sqrt{(l^2 \bigoplus G) - (l \bigoplus G)^2}$, where $\bigoplus$ is the convolution operator and G is a Gaussian kernel with $\mu = 0$ and $\sigma \in \{1.1°, 5.5°\}$.

Additionally, we computed intrinsic dimensionality (Saal, 2010), ID0, ID1, ID2, each with three different kernel sizes (0.12°, 0.52°, 1°), phase-congruency, and phase-symmetry (Kovesi, 1999, 2003) as features. We furthermore considered several interactions of these features. We subtracted red/green contrast, blue/yellow contrast, saturation, and saturation contrast (all finest spatial scale) from phase-congruency and symmetry. Concerning intrinsic dimensionality, we compute ID00.25°– ID01°, ID01° – ID20.52°, ID21° – red/green, ID21°-saturation, ID20.12° – phase congruency. Together with the two last interactions, red-green contrast - saturation contrast and luminance contrast - saturation contrast, this yields 63 different features. Each feature map for each image was z-scored before it was used for further analysis.

To quantify how well a feature can predict fixations and return

locations, we used the area under the receiver-operating characteristics curve (AUC). In short, the AUC assesses how well fixations can be separated from control locations on the same image based on the value of a feature at those locations (Wilming, Betz, Kietzmann, & König, 2011; Tatler et al., 2005). For every feature, we computed the AUC for separating normal fixation locations from control locations and the AUC for separating return locations from control locations. Control locations were chosen from the distribution of fixations on other images, which ensured that control locations follow the spatial distribution of fixations but were not actually fixated locations. We estimated the variability in the data by repeatedly (N=150) computing both AUCs based on 1000 randomly sampled fixation and control locations. Confidence intervals were subsequently bootstrapped (N=2000) on these 150 AUC values for each feature. The dependence between patterns of AUC values was well described by a linear relationship (natural scenes: $r^2 = 0.76$, urban scenes: $r^2 = 0.95$). Figure 6D shows the AUC value of every feature for urban and natural scenes with bootstrapped CIs.

To further investigate the relationship between saliency and return locations, we assigned a saliency score to fixations and return locations. A saliency score was obtained by optimally combining features linearly to separate fixations (or return locations) from control locations. The weights for this combination were estimated with a logistic regression that tried to separate fixations from control locations based upon the 63 features. We used feature values at fixated locations as positive samples and feature values at control locations that were fixated on other-images as negative samples for the logistic regression. To test the hypothesis that return locations are more salient than normal fixations, we estimated two-saliency models and assessed how well return-locations can be predicted in comparison to normal fixations. The two models differ with respect to the samples used for training. The return-saccade (RS) model uses only return locations as positive samples, while the fixation (FIX) model uses only fixation locations from trials where no return

saccade occurred. Both models were trained repeatedly by splitting the available data into test and training sets. We used leave-one-out cross-validation, where each subject was used for testing once and was not used for training in this run, this ensured that training and test data was completely independent. Both models predicted return locations and normal fixations separately. We found that return locations could be predicted with an average AUC of 0.73 (RS: 0.724, FIX: 0.731) compared to an AUC of 0.67 (RS: 0.667, FIX: 0.674) for normal fixations. A two-way analysis of variance with factors 'model type' and 'fixation type' revealed that both main effects and the interaction between the two are significant ($p < 0.0001$)

Fixation Sampling Strategies

To compute an internal priority map for a given subject and image, we computed a 2D histogram of fixation locations of all other subjects that did not make a return saccade on the same image from the same dataset. To obtain a density map, we convolved this histogram with a Gaussian kernel with full-width-half-maximum = 1° and normalized the filtered histogram to unit area.

To evaluate the likelihood that a trajectory is drawn from an internal priority map, we interpreted the internal priority map as cell probabilities for a multinomial distribution. How often a location is fixated gives the counts for each cell. The probability of a trajectory is then given by

$$P(x|m) = \text{multinom}(x_1, ..., x_n; m_1, ...m_n)$$

where x_i encodes the number of fixations for location i, and m_i is the probability of the internal priority map at location i. Subsequently, we compared two different trajectories. In one, the return location is fixated twice, but the last fixation is omitted. In another, the return location is fixated only once, but the last fixation is not omitted. These trajectories differ only in how often the return location and the last fixation are fixated. Thus, the entire comparison amounts to a

comparison of internal priority map values at the return location and the last fixation of the trajectory. However, $P(x_{rs}|m) \geq P(x_{\neg rs}|m)$ is only fulfilled when the priority map value for the return saccade is at least twice as large as the value for the last fixation.

Supplementary Materials

At first glance one might suspect that some remaining structure is present in fixation durations that are corrected with the piecewise linear model (Figure 4E).

However, comparing Figure 4E and Figure 2B (top left) shows that the remaining pattern in 4E is correlated with the number of samples available per bin. We suspected that this and the typically non-normal distribution of fixation durations might introduce statistical effects that could make it likely to observe the remaining pattern.

To estimate the size of such a hypothetical effect, we created a baseline distribution of fixation durations that takes different sample sizes into account by jointly shuffling the angle and amplitude difference labels of corrected fixation durations. We then used this distribution as our null hypothesis that no angle and amplitude difference effect is present and compared the empirical fixation durations against it by bootstrapping (N=1750) confidence intervals for the null hypothesis. At an uncorrected 95% alpha level, we find that in the case of corrected fixation durations, 349 significant deviations (279 expected at p=0.05) occur. Correcting for multiple comparisons leaves only 14 significant deviations.

The significant deviations were spread out uniformly and did not form clusters. Furthermore, as expected, the expected bootstrapped duration per bin is close to zero for every bin. The size of the confidence intervals however varies systematically with the number of bins available per bin (see Figure 3.8).

Because of the above, it is plausible that the remaining pattern in the residuals is due to statistical effects introduced by unequal number of samples per bin and the non-normal distribution of fixation

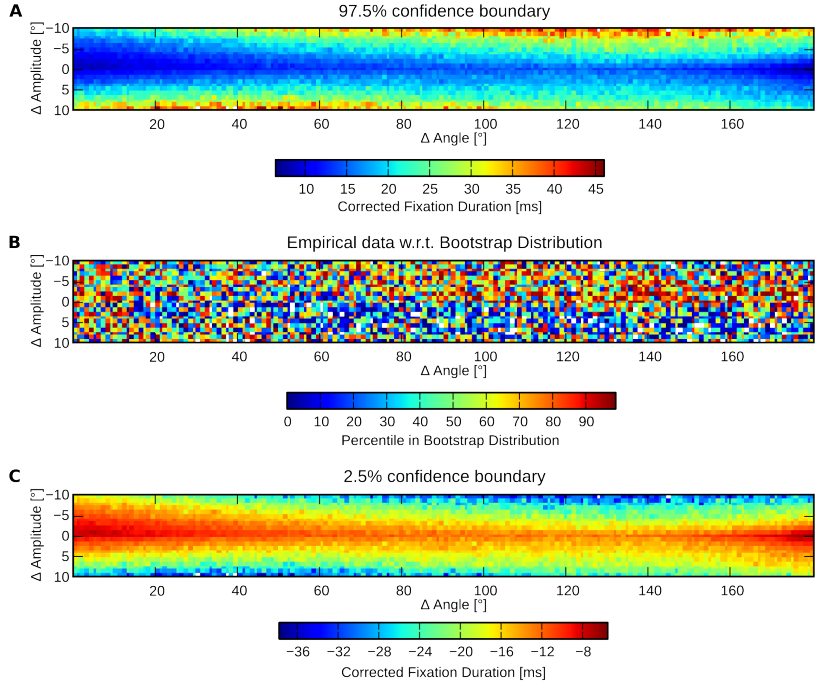


Figure 3.8: Confidence intervals for the hypothesis that no angle and amplitude effect is present in the residuals of the piecewise-linear model. A shows the upper 97.5% confidence boundary as a function of amplitude and angle differences. Values are larger where fewer samples are available. B shows the percentile of the residuals of the piecewise-linear model in the bootstrap distribution. C shows the lower 2.5% confidence boundary. Values are smaller where fewer samples are available.

durations.

Author Contributions

Conceived and designed the study: NW PK. Analyzed the data: NW SH NS PK. Contributed reagents/materials/analysis tools: NW SH NS. Wrote the paper: NW SH NS PK.

Funding

This work was supported by the EU through the project eSMCs (FP7-IST-270212). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors have declared that no competing interests exist.

Chapter 4

Dissociation between saliency signals and activity in early visual cortex

This article has been published in the Journal of Vision: Betz, T., Wilming, N., Bogler, C., Haynes, J.-D., & König, P. (2013). Dissociation between saliency signals and activity in early visual cortex. *Journal of Vision*, *13*(14), 1–12

4.1 Abstract

Saliency is a measure that describes how attention is guided by local stimulus properties. Some hypotheses assign its computation to specific topographically organized areas of early human visual cortex. However, in most stimuli, saliency is correlated with luminance contrast, which in turn is known to correlate with activity in these early areas. Thus, any observed correlation of local activity with saliency might be due to the area encoding luminance contrast. Here we disentangle encoding of local luminance contrast and saliency by using stimuli where the two properties are uncorrelated. First we conduct an eye-tracking study to verify that both negative and positive contrast modifications located in individual quadrants of the visual field increase saliency. Second, subjects view identical stimuli while fMRI BOLD signals are recorded. We find that positive contrast modifications induce a robust increase of activity in V1-V3 and hV4.However, negative contrast modifications lead to a reduced (V1, V2) or comparable (V3, hV4) activity level compared to unmodified quadrants.Furthermore, even with linear multivariate pattern classification techniques it is not possible to decode the location of the salient quadrant independent of the type of the contrast modification. Instead, decoding of the contrast-modified location is only possible separately for the two modification types in V1-V3. These findings suggest that the BOLD activity in V1-V3 is dominated by contrast-dependent processes, and does not include the contrast-invariance necessary for the computation of feature-invariant saliency.

4.2 Introduction

The brain continuously samples information by directing covert or overt attention towards different locations in the environment. In recent years the underlying mechanisms of attentional selection have moved towards the center of research interest. An early hypothesis suggests that the brain computes a saliency map – a topographically

organized representation of the visual field that can be used to decide where to attend next. The more salient a position is, the more likely it will be attended (Koch & Ullman, 1985).

This hypothesis has triggered the search for brain areas that perform the required computations and encode saliency. Several regions in cerebral cortex and subcortical areas have been suggested as the locus of a saliency map: superior colliculus (Kustov & Robinson, 1996), pulvinar (Shipp, 2004), V1 (Li, 1999a; Li, 2002), parietal cortex (Bisley & Goldberg, 2010; Geng & Mangun, 2009; Gottlieb et al., 1998; Serences et al., 2005), V4 (Mazer & Gallant, 2003) and frontal eye fields (Serences & Yantis, 2007; Thompson & Bichot, 2005). The term saliency is often used for task-dependent as well as stimulus-driven processes that might occur in different areas, an ambiguity that may lead to confusion. Here, we are only concerned with the latter. Li Zhaoping (Li, 2002; Zhaoping, 2011) argues that already V1 creates a purely stimulus driven (bottom up) saliency map that relays information to higher areas.

In a network of interacting areas it is difficult to identify and locate computations of saliency. Because V1 projects directly or indirectly to all areas listed above, it is reasonable to assume that saliency information would also be observable in these other areas if it is already computed in V1 (Shipp, 2004; Zhang et al., 2012). Any higher cortical area that receives information from many other areas is therefore likely to exhibit saliency map like properties because saliency related information propagates up the hierarchy. To identify where saliency information is first made explicit, as opposed to simply received, it is therefore important to identify the exact contribution of early visual areas, like V1, to this computation.

A complication arises from the fact that early areas are usually considered to encode certain features of the stimulus (e.g. oriented edges in V1, (Hubel & Wiesel, 1965)), and these features in turn contribute to the saliency map (Itti & Koch, 2001b). Thus, a strong response to a salient stimulus in an area need not reflect the explicit computation of saliency but potentially only the encoding of a feature

that also influences the computation of saliency. For example, high contrast edges strongly activate V1 neurons and high contrast edges correlate with saliency. But encoding of high contrast edges in V1 does not necessarily form an explicit representation of saliency.

In this study, we investigate the contribution of early visual areas of human cortex (V1-V3, human V4) to the computation of visual saliency. We address the dependency of contrast and saliency by exploiting a finding by (Einhäuser & König, 2003): Under certain conditions a local reduction of luminance contrast leads to an increase in saliency. A brain region that explicitly encodes saliency would show an increased activity in response to local contrast reductions (saliency encoding hypothesis). However, a brain region that encodes contrast would show reduced activity (contrast encoding hypothesis). Hence, we created stimuli in which the luminance contrast in one of the four quadrants was either increased or decreased. An eye tracking study confirms that both contrast modifications increased saliency. We then present these stimuli in an fMRI experiment and record BOLD responses. The type of representation is characterized by analyzing the mean BOLD activity and by multivariate pattern classification in functionally defined regions of interest (ROI V1-V3, hV4). This allows differentiating between encoding of luminance and encoding of saliency.

4.3 Methods

Stimuli

We use a set of pink noise images as stimuli to avoid the influence of high level factors and still retain some of the statistics of natural stimuli. We generated these stimuli by randomizing the phases in Fourier transformed natural images, which removes all image structure, but leaves the power spectrum untouched (Einhäuser, Rutishauser, et al., 2006). 27 source images were chosen randomly from the categories natural and manmade (Açik et al., 2010). Their

luminance histograms were flattened, and contrast increases and decreases were applied to each quadrant. Contrast modifications were computed as described before (Einhäuser & König, 2003; Açik et al., 2009). These modifications were chose because they tended to attract fixations in the studies mentioned above. Our stimuli differed from these earlier studies in that the modified area was significantly larger, extending over an entire quadrant. This change was necessary in order to ensure a strong quadrant specific response in the fMRI experiment. Furthermore, the modification includes spatial frequencies sufficiently low to influence saliency. Modified luminance values were given by:

$$I = [1 + \alpha K] * [I0 - M] + M$$

where I0 denotes the source image, M the mean image and α the peak modification level. * is a pointwise multiplication of matrices. α was set to 0.9 for contrast increases and to -0.9 for contrast decreases. I, K and M are matrices of the same size as I0. M contains the local mean luminance values and is computed by convolving I0 with a Gaussian kernel with a full width at half maximum of 6.53°. K describes how the modification level changes as a function of the distance to the peak modification. Here, K is given as the cosine of the square of the distance to the peak modification location.

$$K(x, y) = (cos[\frac{(x2 + y2)}{s}] + 1) * \frac{1}{2}2$$

This function has its maximum value of 1 at x=y=0, and smoothly drops to 0 at $\frac{(x2+y2}{s} = \pi$. s was chosen such that the zero-crossing occurred at the edge of the modified quadrant in the horizontal direction. In the vertical direction, the modification slightly leaks into the adjacent quadrant, but this only corresponds to 0.74% of the kernel's mass. To modify a specific quadrant, K was centered in the respective quadrant. All in all we generated 243 different stimuli (27 stimuli * 2 modifications * 4 quadrants + 27 unmodified; see Figure 1 for examples).

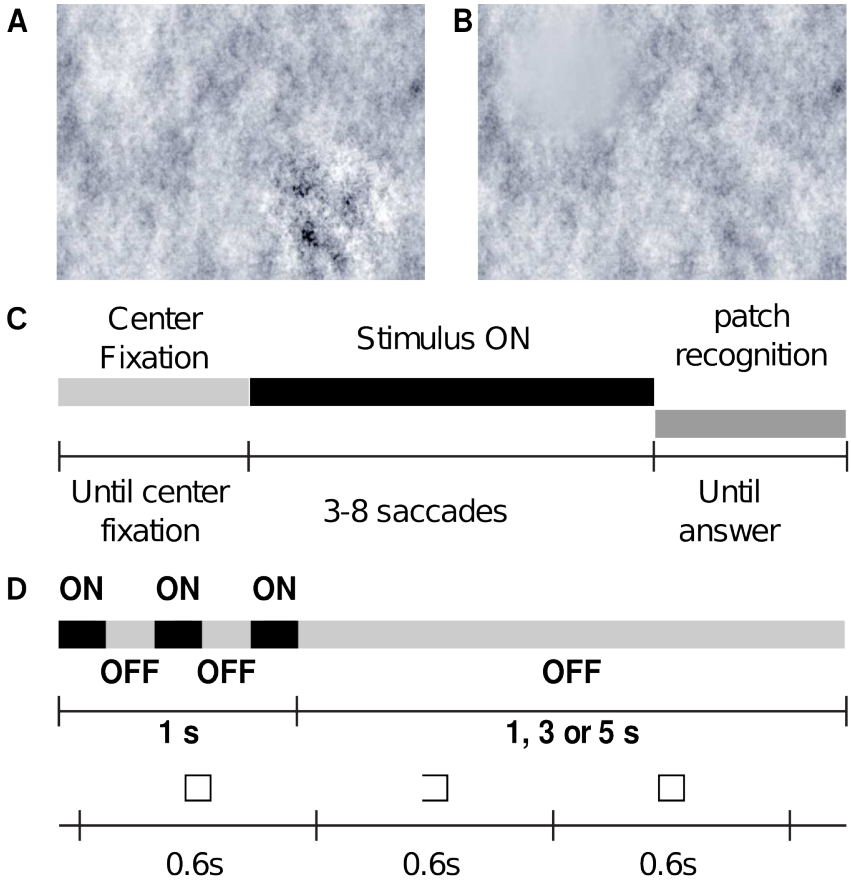


Figure 4.1: Figure 1. A) and B) Examples of pink noise stimuli with high contrast and low contrast modifications. Note that the change in contrast is much more gradual and less visible if stimuli are viewed at their original size. C) Time-course of one eye-tracking trial. Each trial started with a central fixation on a gray screen, after which one pink-noise stimulus was shown until 3-8 saccades were completed. In 49 out of 243 trials a patch recognition task was performed after stimulus offset. D) Time-course of an fMRI trial. In each trial, one image was presented repeatedly for 200 ms with a 200 ms gap. During a gap the screen switched to a gray background. Between successive trials there was a variable interstimulus interval of 1 to 5 s. Observers had to detect the opening of one side of a central square. This task ran continuously throughout the session, and independently of the pink-noise stimulation.

Eyetracking

Participants

Eleven student volunteers took part in the eye tracking experiment (4 male and 7 female, age range 20- 30 years, mean age 25 years). All participants had normal or corrected to normal vision. Inclusion in the study was contingent on reliable eye tracking calibration with an average validation error below 0.3°. As compensation, participants received either payment or course credit. The study was conducted in accordance with the declaration of Helsinki, and approved by the local ethics committee.

Apparatus

The experimental apparatus was designed to resemble the parameters of the fMRI experiment. Screen distance was 80 cm to achieve the same stimulus size as later on in the scanner (26.6° x 20.5°). Stimuli were presented on a 19-inch flat screen monitor (SyncMaster 971p, Samsung Electronics, Seoul, South Korea) at a screen resolution of 1280x1024 pixel and refresh rate of 75Hz. Eye movements were recorded with an Eyelink II system (SR Research Ltd., Mississuaga, Ontario, Canada) at 500Hz sampling rate. The system is capable of tracking both eyes, but only the eye that gave a lower validation error after calibration was recorded. A chin-rest was used to minimize head movements. The experiment was conducted in a darkened room.

Procedure

The stimuli were presented in 3 blocks of equal length. In the break between blocks we encouraged participants to rest and remove the eye tracker. Before each stimulus onset drift correction was performed, requiring participants to fixate the center of the screen. Subsequently, each stimulus was presented until a random number of saccades between three and eight had been performed (see Figure

1C). Each stimulus was presented once to each subject (243 trials per subject, 2673 trials in total). The stimulus order was randomized, but the number of stimuli from each condition was the same in all blocks. The task was to recognize whether a patch of size 250x250 pixels, presented after stimulus offset, was taken from the image just shown. The probability that the patch actually came from the previously seen image was 50%. Participants responded by pressing either the arrow up button for yes or the arrow down button for no on a regular keyboard. To shorten the duration of the experiment and to avoid fatigue or de-motivation, the patch recognition task was only presented after 49 randomly selected trials out of the 243. A test run, consisting of five images, was performed in order to let the subjects gain experience with the task.

fMRI

Participants

14 naive observers participated in the fMRI study (8 female, 6 male; age range 22-33 years, mean age 27 years). They reported normal or corrected to normal vision, and received payment for their participation. The study was conducted in accordance with the declaration of Helsinki, and approved by the local ethics committee. For all participants, detailed anatomical brain images were available from previous studies. The data of two observers had to be excluded from analysis. One fell asleep during the experiment, and for the other, retinotopic mapping was not successful.

Apparatus

Images were presented with a Sanyo Xtra Pro (SANYO Electric Co., Ltd., Osaka, Japan) with a resolution of 1024x768 pixel. Only the central 800x600 pixel (26°x20°) were in clear view, so images were resized to this resolution with bicubic interpolation. A Siemens 3T Magnetom (Siemens AG, Erlangen, Germany) was used to acquire

functional MR EPI volumes with 36 slices at an isotropic resolution of 3 mm3 (TR = 2000ms; TE=24ms; 36 axial slices; FOV 192x192x108 mm; $\alpha = 0°$). Structural images were acquired with a T1-weighted 3D MP-RAGE with selective water excitation and linear phase encoding. Magnetization preparation consisted of a nonselective inversion pulse. The imaging parameters were TI = 650 ms, TR = 1300 ms, TE = 3.93 ms, $\alpha = 10°$, spatial resolution 1 mm3 isotropic, two averages.

Procedure

The pink noise experiment was divided into five different runs of stimulus presentations. Between each run, participants could take a small break and relax their eyes. Each run lasted approximately 10 minutes and consisted of 108 presentations of pink noise stimuli. Each run contained an equal number of stimuli from each condition, interleaved with 30 blank trials. In contrast to the eye tracking experiment, a task was given that required fixating the center of the screen. The outline of a 0.3°x0.3° square was drawn in black in the middle of the screen. Every 1200ms either the left or right side of the square opened for 600ms and participants had to indicate which side was opened by pressing one of two buttons with the middle or index finger of their right hand. The fixation task was independent of the stimulus presentation and ran continuously during each run. Each run started with ten seconds of the fixation task without pink-noise stimulation. Each stimulus presentation consisted of one pink noise stimulus being flashed three times for 200ms in the background, with a 200ms gap between flashes. Between flashes the background was gray. The rationale for flashing the stimulus was that we wanted to mimic the condition of the eye-tracking experiment just before the target of the first free saccade is chosen. There the stimulus appears with a 'flash' relative to the gray background of the drift correction screen. We chose to flash the stimulus several times to ensure that the pink-noise stimulus was not completely ignored despite being irrelevant for the fixation task. Additionally this procedure was used in a previous study where saliency processing effects could be

decoded successfully (Bogler, Bode, & Haynes, 2011). The inter-trial interval was variable, 1s, 3s or 5s (see Figure 1C). 286 functional MRI volumes were acquired in each run. A 42-slice whole brain EPI image was also acquired to facilitate spatial normalization.

Retinotopic mapping and localization runs were conducted in a separate session on a different day. The retinotopic mapping runs consisted of the presentation of a rotating wedge (5cyc/300s) and an expanding ring (10expansions/300s). This allowed for a functional definition of early visual areas, especially V1-V3 and hV4 (Wandell, Dumoulin, & Brewer, 2007; Warnking, 2002). To localize voxels that react to visual stimulation of a quadrant, flickering checkerboard patterns that were centered in one of the four quadrants were shown (quadrant localizer). These decreased in contrast according to the same spatial function that was used for the contrast modification in the pink noise stimuli. One localizer run consisted of the sequential presentation of localizer images for all quadrants. Each image was presented for 7.5s and changed polarity with 10Hz. The order of stimulation was: upper left, upper right, lower left and lower right and was repeated 10 times. All in all, the mapping and localization session consisted of eight different runs in the following order: 2x rotating wedge, expanding ring, 2x rotating wedge, expanding ring, 2x quadrant localizer. 155 volumes were acquired in each run, but no stimulation was present during the last 10s (five volumes). Participants had the same fixation task as in the pink noise experiment, with the only difference that the fixation spot changed every 1000ms.

Data processing

Functional brain scans were pre-processed with SPM2 (http://www.fil.ion.ucl.ac.uk/spm). The first five volumes of each experimental run were discarded to allow for magnetic relaxation effects. All volumes acquired during one experimental session were motion corrected, realigned to the initial scan of the experiment, and coregistered to the high resolution anatomic image of the participant. For subsequent statistical analyses of the pink noise experiment, a

general linear model (GLM) with event-based and HRF-convolved regressors was estimated separately for each voxel. For every run 9 regressors were used that encoded stimulation onsets: one regressor for no modification, four regressors for a high contrast modification in one of the quadrants and four regressors for a low contrast modification in one of the quadrants. In addition, a constant regressor for each run was included. All analyses were carried out based on SPM parameter estimates for these regressors.

Definition of functional regions of interest

Visual areas V1-V3 and hV4 were functionally defined using well-established retinotopic mapping procedures (Wandell, Dumoulin, & Brewer, 2007; Warnking, 2002) First, we segmented gray matter using FreeSurfer (Dale, Fischl, & Sereno, 1999). Next, the cortical surface was flattened with mrGray (Wandell, Chial, & Backus, 2000). Custom Matlab (The MathWorks, Natick, MA) scripts were used to generate the flattened angular phase maps (Heinzle, Kahnt, & Haynes, 2011). Finally, we identified visual areas V1-V3 and hV4 by locating phase reversal boundaries on these maps. The precise definition of V4 in humans is still debated (Goddard, Mannion, McDonald, Solomon, & Clifford, 2011). Here we use the definition proposed by (Wandell et al., 2007). Figure 2 shows the outlines of these areas on a flattened cortex for one participant and one hemisphere. Note that some smoothing in Figure 2 is due to the visualization and was not present when determining ROI boundaries.

The quadrants of the visual field were localized by fitting a GLM with one event-based and HRF-convolved regressor for each quadrant stimulation onset. Quadrant ROIs contained voxels that were exclusively active when one particular quadrant was stimulated (t-test $p<0.001$ uncorrected), but not during stimulation in one of the other three quadrants. Figure 2 shows outlines of quadrant specific regions for one participant and one hemisphere. Note that voxels which showed specificity for more than one quadrant were discarded. Across subjects 44, 44, 37 and 31% (V1,V2,V3, hV4) of the

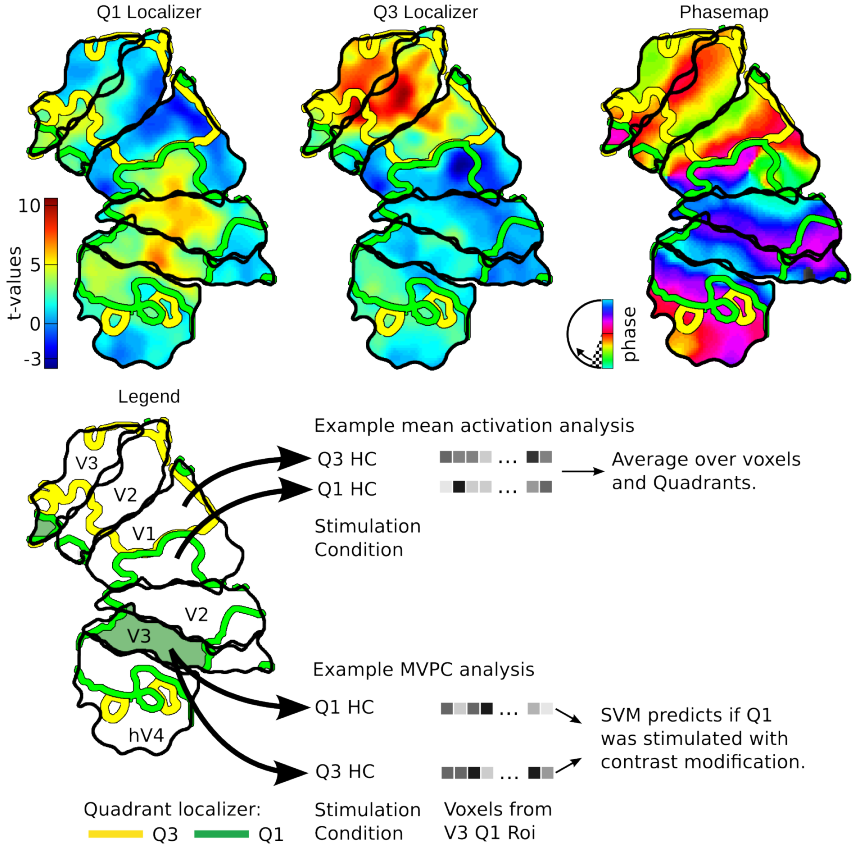


Figure 4.2: Top row: Top left and top center panels show t-maps of one observer for stimulation of a quadrant (Q1 left, Q3 center) with a flickering checkerboard. The colored outlines mark areas that show significant activation to stimulation of Q1 (green) or Q3 (yellow), and no significant activation to stimulation of any other quadrant. The black outlines show early visual areas identified by retinotopic mapping on a flattened cortex for one observer. The phasemap used for identifying early visual areas by locating phase reversal boundaries is shown on the top right. Bottom row: Depiction of mean activation analysis and the multivoxel pattern analysis. The shaded green area highlights as an example the ROI V3 Q1. The two analyses were carried out for each visual area (mean activation analysis) and quadrant area combination (multivoxel pattern analysis).

voxels that showed selectivity for some quadrant were selective for only that quadrant.

Multivoxel pattern analysis (MVPA)

The contrast encoding hypothesis predicts that early areas show increasing activation with increasing contrast. This implies that a classifier can distinguish if a set of quadrant specific voxels was stimulated with low, baseline or high contrast. The saliency encoding hypothesis predicts that the neural response to stimulation with equally salient low and high contrast modifications is comparable. This implies that a classifier can distinguish if a quadrant was stimulated with a low or high contrast modification, even if the training data does not contain information about the modification type.

To test these predictions, we trained support vector machines (SVM) for classifying whether a quadrant received a contrast modification. The SVMs were trained to predict, based on the activation of voxels in one ROI (e.g. V1 Q1), whether the quadrant corresponding to the ROI (Q1) or another quadrant was stimulated with modified contrast. This implicitly compares activation within a ROI when it is stimulated with a contrast modification with the activation when it is stimulated with baseline contrast. We conducted three separate classification analyses: First, a high contrast classifier that only received training and test data from conditions with a contrast increase in one quadrant; second, a low contrast classifier that only received low contrast stimulation parameter estimates; third, a saliency classifier received data from both conditions (twice as many individual data points as the modification-specific classifiers), based on the rationale that neurons encoding saliency should show similar responses to both modifications.

SVM classification for each region of interest was based on three pairwise comparisons, separating contrast modification in the quadrant corresponding to the ROI (=stimulated) from modification in one of the other quadrants (=not stimulated). For example, the high contrast classification accuracy for the ROI V1 Q1 would be the mean

of three accuracies: high contrast (hc) in Q1 vs. hc in Q2; hc in Q1 vs. hc in Q3; and hc in Q1 vs. hc in Q4. Training and evaluation of the SVMs was performed in a leave-one-out cross validation scheme. In each cross validation step, SPM model parameter estimates from 4 of the 5 experimental runs were used to train a classifier that predicted the location of the modified quadrant in the 5th run. This procedure was carried out for each participant and each ROI individually. The cost parameter was set to one.

We also performed an analysis of the separation hyperplanes created by the SVMs in order to gain a better understanding of the representation of contrast in the different visual areas. If the neural response to stimulation with low and high contrast is the same, the SVMs should learn the same separating hyperplane for comparing low contrast vs. baseline and high contrast vs. baseline. The saliency encoding hypothesis therefore predicts that the normal vector to the separating hyperplane points in the same direction for comparing low-contrast vs. baseline and high-contrast vs. baseline. Conversely, the contrast encoding hypothesis predicts that the two normal vectors point in opposite directions. This is illustrated geometrically in Figure 5B. We therefore computed the weight vectors (the normal vector to the separating hyperplane) for pairs of SVMs that predicted low contrast vs. baseline and high-contrast vs. baseline stimulation based on data from the same runs. These weights were then averaged over the five runs and the three different non-ROI quadrants. In a next step we computed the angle between those weight vectors. If the two weight vectors are completely independent, for example if weight vectors are not consistent across runs of the experiment, the expected angular value is 90°. Angles significantly above 90° indicate that the contrast response is greater than a potential saliency response and angles significantly below 90° indicate a stronger saliency than contrast response.

4.4 Results

High and low contrast modifications increase saliency

The primary goal of our study was to disentangle computations of luminance contrast and saliency. We created images on which luminance contrast in a quadrant was either decreased or increased. The attentional effect of these modifications was first investigated in an eye-tracking study. We recorded how often the first free fixation on an image fell into each quadrant. Analysis was restricted to the first fixation because its target is selected while the retinal stimulation is identical to the central fixation in the fMRI task. The distribution of fixations across the different quadrants is shown in Figure 3A. Each quadrant attracts more fixations in each of the two modification conditions than when it is unmodified (Fig. 3B). This is backed by a two factorial repeated measures analysis of variance with quadrant and modification as factors. Only the modification factor is significant (modification $p<0.001$; quadrant $p>0.3$; interaction $p>0.5$). Importantly, both modifications are significantly different from baseline (high contrast vs. baseline $p<0.001$; baseline vs. low contrast $p<0.001$, t-test). We conclude that both increases and decreases in local luminance contrast increase saliency in the modified quadrant by a comparable amount. Thus, these stimuli are suitable for disambiguating between the retinotopic processing of luminance contrast and saliency.

Mean BOLD activity increases with contrast

We analyzed how contrast modifications affected the mean BOLD response to the modified image regions in brain regions that process the visual input. We extracted GLM parameter estimates from all voxels in 16 functionally defined regions of interest (ROI) corresponding to the four quadrants of the visual field in V1-V3 and hV4 (see method section and Figure 2 'Example mean activity analysis'). The contrast encoding hypothesis predicts low activity in

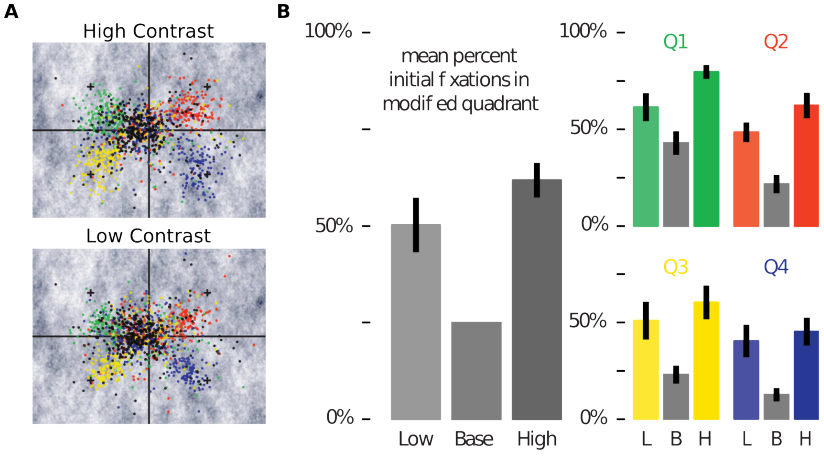


Figure 4.3: Distribution of fixations in the different stimulus conditions. A) Each color encodes fixations made when a certain quadrant was modified (Q1: green; Q2: red; Q3: yellow; Q4: blue). Gray fixations were made on unmodified stimuli. Solid gray lines mark the quadrant borders, plus-signs mark the peaks of the modification which spanned the entire quadrant. Neither were shown on the actual stimuli. For all modifications, the fixation distribution is shifted towards the peak of the modification. B) Mean ratio of fixations made in each quadrant. Small colored figures show data for individual quadrants, the larger gray diagram shows the mean across all quadrants. Errorbars indicate the standard error of the mean across subjects. All quadrants attract more fixations when they are modified than in the baseline condition. This effect is independent of the direction of the contrast modification.

quadrants stimulated with reduced contrast, and high activity for high contrast stimulation. The saliency encoding hypothesis, in its strongest form, predicts increased activity for both types of modification compared to baseline. We analyzed activity averaged across quadrants in individual areas in the high contrast condition, low contrast condition, and for the unmodified images (Figure 4). A repeated measures ANOVA with condition (high contrast, baseline, low contrast) and area (V1-V4) as factors reveals significant main effects of both factors ($p<0.001$), and a significant interaction ($p<0.05$). Single factor ANOVAs computed on the data of individual areas show that the effect of condition is significant throughout all areas ($p<0.001$ Holm-Bonferroni corrected). We assessed the source of the significant effect with post hoc pairwise t-tests. The differences between high contrast and baseline as well as between high contrast and low contrast are significant in all areas ($p<0.01$). The difference between low contrast and baseline, the latter inducing higher activity than the former, is only significant in V1 ($p<0.01$) and V2 ($p<0.05$, all values Holm-Bonferroni corrected). In summary, high contrast leads to an increase in activity compared to both baseline and low contrast condition, but low contrast, although salient, does not likewise lead to increased activity. To the contrary, if there is any difference between low contrast and baseline condition, it is not in the direction predicted by the saliency processing hypothesis.

MVPA supports contrast encoding hypothesis

In principle, it is possible that neurons in V1-V4 encode saliency, but that this information is represented in these areas in a way not accessible to an analysis of the activity level in the form of an averaged BOLD response. We used multivoxel pattern analysis (Haynes & Rees, 2006; Kriegeskorte, Goebel, & Bandettini, 2006) to test if information about the most salient quadrant can be decoded from the activity patterns in our ROIs. The contrast encoding hypothesis predicts that the stimulation of a visual field quadrant with a certain level of contrast leads to a specific pattern of activity in the ROI

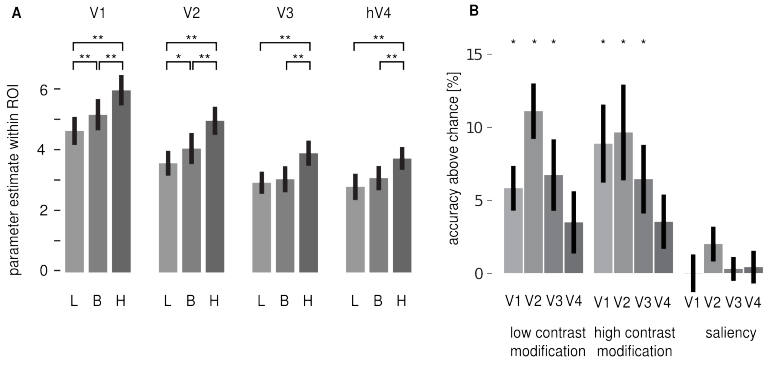


Figure 4.4: A) Mean BOLD activation in different visual areas in the 3 contrast conditions (L = low, B = baseline, H = High) averaged across quadrants. In all areas, an increase in contrast leads to either no change or an increase in BOLD signal, but never a decrease. Errorbars represent standard errors of the mean across subjects. Asterisks indicate significant differences between conditions (pairwise t-tests, Holm-Bonferroni corrected; *: $p<0.05$, **: $p<0.01$). B) Mean decoding accuracies above chance level (50%) for linear SVMs trained to predict whether a quadrant received a given modification. Errorbars represent SEM across subjects, asterisks indicate prediction performance significantly above chance assessed by a t-test ($p<0.05$).

corresponding to that quadrant. It should thus be possible to decode whether the stimulus was modified in the quadrant corresponding to a ROI (see Figure 2). For example, given an activation pattern from ROI V3 Quadrant 1 (Q1), induced by high contrast stimulation in either Q1 or a different quadrant, it should be possible to infer if Q1 or another quadrant was modified (see Figure 2 'Example MVPC analysis'). The same should hold for low contrast modifications. The saliency encoding hypothesis furthermore predicts similar activation patterns for low and high contrast modifications, since both make a quadrant more salient (Figure 3). A classifier trained on both types of patterns combined should therefore be able to generalize, and infer if a quadrant was modified even without knowledge of the modification type. Figure 4B shows the mean decoding accuracies above chance level (50%) achieved for the 3 different analyses (high contrast only, low contrast only, both contrasts mixed = saliency) in V1-V3 and hV4. In areas V1 through V3, decoding accuracies were significantly above chance level for the high contrast only and low contrast only analyses (t-test across 12 subjects, $p<0.05$). However, the decoding accuracy for the saliency analysis did not reach significance in any ROI. Since the difference between a significant result and a non-significant one is not necessarily itself significant (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011), we also directly analyzed the differences in accuracy between the contrast classifiers and the saliency classifier. Here, we find that in areas V1 through V3, decoding accuracy is significantly higher for the contrast classifiers. Results for hV4 are not significant, but the trend goes in the direction predicted by the contrast encoding hypothesis.

The analysis of the weights of the classifiers shows that on average, voxels have a positive weight for high contrast, and a negative weight for low contrast (see Figure 5A; V1 9 out of 12 subjects in lower right quadrant, V2 8 out of 12, V3 7 out of 12, hV4 6 out of 12). For areas V1-V3, the normal vectors for the two hyperplanes (high contrast within ROI vs high contrast outside ROI and low contrast within

ROI vs low contrast outside ROI) tend to point in opposite directions, i.e. the angle between these vectors is significantly greater than 90° (Figure 5C, individual t-test across 12 subjects, $p<0.05$). All of these results are expected and corroborate the univariate analysis.

To ensure that the failure to decode saliency from early visual areas is not due to specific parameters chosen, we performed additional analyses. First, we trained SVMs on activation patterns from whole areas, instead of only single quadrants. And second, we used anatomical ROI definitions from the SPM anatomy toolbox (Amunts, Malikovic, Mohlberg, Schormann, & Zilles, 2000; Eickhoff et al., 2005; Rottschy et al., 2007) instead of the functional ones. Neither of these changes, nor combinations thereof, affected the pattern of results reported above.

In summary, it is possible to decode whether a quadrant was modified when modification type is given. However, without this information it is not possible to decode whether a quadrant is salient. This suggests that V1-V3 and hV4 do not make the abstraction away from absolute changes in contrast to changes in saliency.

4.5 Discussion

We showed that low and high contrast modifications in pink noise stimuli decouple saliency and contrast. Our eye tracking data indicate that both types of modification increase saliency. This decoupling provides a tool for investigating saliency processing in fMRI BOLD responses in early topographically organized visual areas (V1-V3, hV4). The behavioral increase in saliency for the low contrast modifications is not mirrored in fMRI data. Instead, we found that the activity patterns of V1-V3 monotonically relate to stimulus contrast, not saliency.

In order to encode saliency for these stimuli, the visual system would have to increase its response to contrast deviations from the mean in both directions.(Gardner et al., 2005) have shown that such a rectification operation may happen in hV4 during temporal contrast

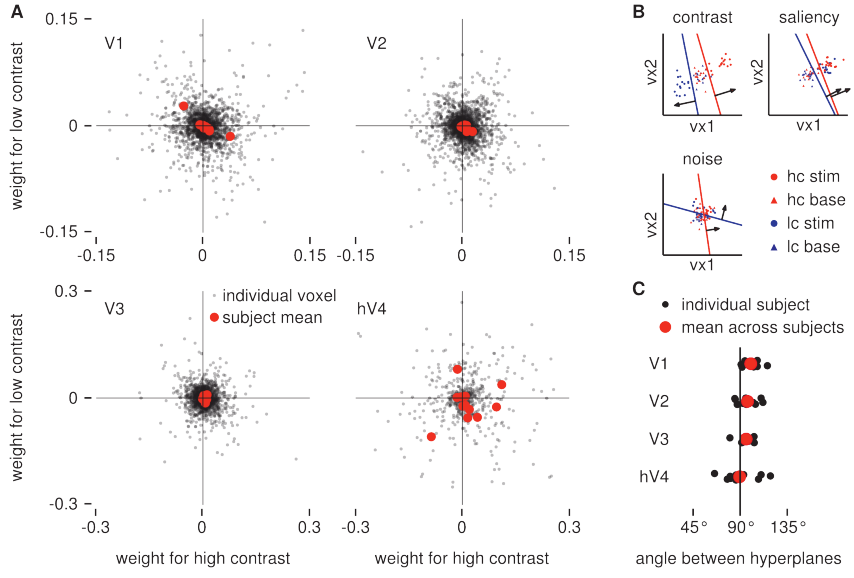


Figure 4.5: A) Linear SVM weights assigned to individual voxels for high contrast and low contrast modifications. Black circles mark individual voxels of all subjects, red circles mark the mean of a subject. Mean values tend to cluster in the lower right quadrant, indicating that voxels received a positive weight for the high contrast condition and a negative weight for the low contrast condition. (V1 9 out of 12 subjects in lower right quadrant, V2 8 out of 12, V3 7 out of 12, hV4 6 out of 12). B) Illustration of the rationale behind the analysis of angles between separating hyperplanes. If two voxels encode contrast, low contrast stimulation will lead to lower activity then baseline stimulation, which in turn leads to lower activity the high contrast stimulation. The normal vectors to the hyperplanes separating low contrast from baseline and high contrast from baseline point in opposite directions. If the voxels encode saliency, low contrast stimulation also leads to higher activity than baseline, and both normal vectors point in the same direction. If there is no difference between the stimulation conditions (labeled noise here), the normal vectors are uncorrelated, and on average the angle between them will be 90°. C) Angles between hyperplanes in the four visual areas. Black circles mark individual subjects, red circles indicate the mean. In areas V1-V3, the mean is shifted right of the 90° line.

adaptation. It might have been suspected that a similar mechanism for spatial variations in contrast is responsible for the behavioral saliency effect observed in our stimuli. We do not find evidence for this. Low contrast stimulation did not lead to an increased activity level compared to baseline and saliency could not be decoded in hV4. However, contrast could also not be decoded in hV4, which might be indicative of a low signal to noise ratio. The question of whether saliency is encoded in hV4 can therefore not be conclusively addressed with our data.

The V1 saliency hypothesis (Li, 2002) states that activity in V1 creates a bottom-up saliency map. Specifically, the highest evoked V1 response of each visual field location (i.e. a max operation over all features encoded for this location) gives the relative saliency of this location. There is psychophysical (Zhaoping & May, 2007) as well as physiological (Zhang et al., 2012) evidence supporting this hypothesis. At first sight these data appear to be in conflict with the present results. However, their stimuli are not natural stimuli, but arrays of oriented bars or simple conjunctions of bars. It is known that the receptive fields of V1 neurons are highly tuned to such bars. Under these conditions it is therefore plausible that processing in V1 contributes to a saliency map. Given the restricted stimulus set, focusing on oriented line elements, the intermediate results might be indistinguishable from the final saliency map. In contrast, we used stimuli with a power spectrum that is comparable to natural scenes. Recent work demonstrates that such stimuli induce qualitatively different dynamics in visual cortex than gratings (Onat, König, & Jancke, 2011). Hence, our more complex stimuli might explain why we find that V1 BOLD activity only contributes one processing step on the way to a final saliency map. This is consistent with recent results on experimental blindsight in monkeys (Yoshida et al., 2012). Interestingly, there is even evidence for salient orientation pop-out stimuli which are represented in V4 rather than in V1 (Bogler et al., 2011). These results are not compatible with the predictions of a general saliency map localized in primary visual cortex.

It should be noted that our results do not rule out contributions of V1-V4 to the computation of saliency even in the low-contrast modification condition. It might, for example, be that subpopulations of neurons in these areas compute saliency and that the activity of these subpopulations is swamped by the contrast dependent activity changes of the majority of neurons. However, Zhang et al., 2012 do find an explicit attention driven signal in BOLD responses in V1 regions even for stimuli that were not consciously perceived. Thus, it does not seem that the proposed V1 saliency map is in principle not discoverable with fMRI. This concern is further reduced by our use of decoding techniques. It has been shown that MVPC analyses can be successfully used to decode the activity of neuronal subpopulations below the spatial resolution of individual fMRI voxels (Haynes & Rees, 2006). But this is of course still no guarantee that decoding would have been successful in our case if saliency is encoded by a small set of neurons in early visual areas. However, more explicit representations of saliency are observable in higher visual areas with fMRI (Bogler et al., 2011). Summarizing, the most dominant feature in V1-V3, according to our analysis, is clearly luminance contrast and not saliency.

In conclusion, we report a case of behaviorally observable saliency that is not linearly driven by stimulus contrast. Our findings do not support the hypothesis that a saliency map, in the sense of an explicit representation of most likely fixation target regardless of specific stimulus features, is found in V1-V3. It is conceivable that higher areas have to integrate feature specific saliency information encoded in early processing stages.

This work was supported by two grants by the German Federal Ministry of Education and Research (BMBF grant 01GQ1001C, BMBF grant 01GQ0851), by the EU through the project eSMCs (FP7-IST-270212) and ERC-2010-AdG #269716 - MULTISENSE, the Deutsche Forschungsgemeinschaft (GRK1589/1), and the Max Planck Society.

Chapter 5

Differential contribution of low and high-level image content to eye movements in monkeys and humans.

This chapter is joint work with Megan Jutras, Elizabeth Buffalo and Peter König:
Wilming, N., Jutras, M., Buffalo, E. A., & König, P. (n.d.). Differential contribution of low and high-level image content to eye movements in monkeys and humans. *(In Preparation)*

Abstract

"The eyes are a window to the soul" is an old proverb that suggests that our eye-movements reveal aspects of our mind that are otherwise private. What we look at, and how we select where to look, therefore offer unique opportunities to understand the mind. Consequently, comparing eye-movement behavior between humans and macaque monkeys allows to better understand our most prominent model for investigating the neural basis of the human mind. Here we are interested in whether the selection of gaze locations is comparable in macaque monkeys and humans. We investigate whether local stimulus features and more complex stimulus content influence gaze location selection to the same extent in both species. First, we show that a stimulus dependent salience, model that only considers local image features, predicts monkey fixation locations. Crucially, we find that the same pattern of local image features predicts fixation locations in both species to a comparable extent. This is compatible with the common view that humans and monkeys share a neural system for guiding eye-movements. Second, human gaze locations are very well predicted by locations marked as "interesting" by independent observers. Importantly, the stimulus dependent salience model predicts monkey eye-movement behavior better than interesting locations. Intuitively this suggests that human, but not monkey, eye-movements are guided by a common understanding of the semantic stimulus content. However, the same pattern of results was observed during the viewing of fractal images which are devoid of semantic content. In summary, our results suggests that humans and monkeys share a common salience system, but that picture viewing activates additional stimulus- dependent processing resources in humans that are either absent or not activated in monkeys.

5.1 Introduction

During action and perception in their natural environment humans direct their eyes towards relevant regions of the visual space. Due to the sharp drop of visual acuity from the central visual field, i.e. at the fovea, to higher eccentricity such fixational eye movements have a crucial influence on which part of potential sensory information is selected for further processing. Understanding the neural basis of eye-movement target selection is therefore fundamental for understanding human cognition at a larger scale (Petersen & Posner, 2012).

Understanding control of eye movements in humans requires detailed anatomical, physiological and behavioral studies. However, for obvious reasons, not all currently available techniques are suitable for human studies. Specifically, invasive recordings with microelectrodes in monkeys have generated a vast amount of knowledge. The macaque monkey is widely considered to be the model system for studying the neural basis of overt attention and oculomotor control (Bisley, 2011). Yet, only few studies have investigated systematic similarities and differences between human and monkey selection of fixations under free viewing conditions (Berg, Boehnke, Marino, Munoz, & Itti, 2009; Einhäuser, Kruse, Hoffmann, & König, 2006; Kano & Tomonaga, 2009; Kayser, Nielsen, & Logothetis, 2006; Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010; McFarland et al., 2013). Remarkably, the grand total of monkey subjects in these studies adds up to only 18. Thus, our (limited) knowledge of eye-movement behavior in monkeys under natural conditions rests on a very small database. In order to generalize insights obtained by studies on the monkey visual and oculomotor system comparative studies are needed.

The study of eye-movement behavior in humans and monkeys covers a large variety of paradigms. These range from tightly controlled settings under laboratory conditions (e.g. fixate for an extended period of time and perform a specific eye movement after the go signal)

to free viewing of images (e.g. look wherever you like). In daily life the task may exert a substantial influence on the selection of fixation point (Yarbus, 1967; Land & Tatler, 2001; Rothkopf et al., 2007; Ehinger et al., 2009; Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010; Betz et al., 2010; Sullivan et al., 2012; Johnson et al., 2014). However tasks are typically specified on an abstract level. Recently investigated tasks range from making a sandwich (Johnson et al., 2014), to the search for pedestrians (Ehinger et al., 2009). Up to now it has proved difficult to parametrize and compare such tasks. And indeed, in the investigation of eye movements while scrutinizing complex stimuli, free viewing is still the dominating "task". It has been argued that "free viewing" is not equivalent to "no task", but just states the ignorance of the experimenter regarding the agenda selected by the observers (Tatler et al., 2011). From a Bayesian perspective, this amounts to sampling tasks from the prior distribution of tasks. We believe that this is a reasonable definition for the absence of a specific task. Free viewing furthermore allows humans and monkeys to select the tasks that are preferable for them while scrutinizing pictures. Free viewing is therefore a suitable baseline for comparing natural viewing behavior between species.

Eye movement behavior during free viewing is thought to vary along several dimensions. An important distinction is made between stimulus-dependent, context dependent and geometrical aspects in the guidance of eye movements. Stimulus-dependent influences on viewing behavior are, for example, the saliency conveyed by low-level images features (Itti & Koch, 2001b; Parkhurst et al., 2002; Peters et al., 2005), higher-level influences like objects (Einhäuser et al., 2008; Nuthmann & Henderson, 2010) and scene interpretation (Kietzmann et al., 2011). Context dependent aspects include the task (Betz et al., 2010; Castelhano, Mack, & Henderson, 2009) and context (Kietzmann et al., 2011; Torralba et al., 2006). Oculomotor biases like the center bias of fixations (Tatler & Vincent, 2009) or saccadic momentum (Smith & Henderson, 2011b; Wilming et al., 2013) consistently make strong contributions (Kollmorgen et al.,

2010) to the guidance of eye movements. In the present article we focus on stimulus-dependent guidance of eye movements in humans and monkeys.

Salience conveyed by low-level image features is an important component of stimulus dependent guidance. Low level image features are thought to mimic results of processing in the early visual system which are transmitted to more central structures (Li & Zhaoping, 2005; Bogler et al., 2011; Soltani & Koch, 2010). Features are subsequently weighted and combined into a single map that encodes salience of potential targets for fixations. This process is known under the name of bottom-up salience and we follow this convention here. Whether or not such a salience map exists in the human brain, and if yes where, is a matter of ongoing debate (Mazer & Gallant, 2003; Li & Zhaoping, 2005; Arcizet, Mirpour, & Bisley, 2011b; Bogler et al., 2011; Betz et al., 2013). In many cases salience models have been successfully used to predict eye-movement targets during free viewing of images (Bruce & Tsotsos, 2009; Einhäuser et al., 2008; Hwang et al., 2009; Itti & Baldi, 2005b; Judd, Ehinger, Durand, & Torralba, 2009; Kienzle et al., 2007; Einhäuser & König, 2003; Land & Hayhoe, 2001; Zhang et al., 2008; Zhao & Koch, 2011). Furthermore, recent reports provide evidence for the existence of a functional saliency map in the human brain (Ossandón et al., 2012).

The contribution of salience based on low-level visual features for the selection of fixation points in monkeys has been investigated by a small number of groups. These consistently report above chance predictions of fixation points by salience models in both humans and monkeys for videos clips (Berg et al., 2009) and gray scale images (Einhäuser, Kruse, et al., 2006). But there is considerable disagreement in how far the effects are qualitatively comparable in humans and monkeys. Einhäuser, Kruse, et al. (2006) argue that both species are equally driven by bottom-up salience, while Berg et al. (2009) report that bottom-up salience is more predictive for human eye movement behavior. Notably in both cases the same salience model was used for humans and monkeys. This makes interpretation of

the results difficult because bottom-up salience might be computationally different but equally effective in both species. For example, humans and monkeys might emphasize different low-level properties, such as luminance contrast or corners, of the stimulus but the mechanism for selecting salient regions from low-level image properties might be the same. Thus, identifying how, and how strongly, bottom-up salience guides eye movements in monkeys and humans is crucial to establishing whether monkeys are a good model system for bottom-up saliency driven eye movement behavior in humans.

In humans, bottom-up salience contributes to the guidance of eye movements consistently, but only to a limited extent (Schütz, Braun, & Gegenfurtner, 2011; Kollmorgen et al., 2010). But exactly what other factors are relevant during free viewing of scenes is an ongoing matter of debate. Similarities between monkeys and humans exist for example during the viewing of faces (Ghazanfar, Nielsen, & Logothetis, 2006; Shepherd et al., 2010) and viewing of scenes containing simple social interactions (McFarland et al., 2013). While these findings are important, they do not assert that viewing behavior between species is similar when other higher-level factors guide eye-movements. Recently several groups circumvented the problem of identifying higher-level features by asking independent observers to mark "interesting" locations in images (Elazary & Itti, 2008; Masciocchi, Mihalas, Parkhurst, & Nierbur, 2009; Onat et al., 2014). They found that "interestingness" was by a good margin the best predictor of human eye movement behavior. Thus, to determine if monkeys might serve as a model system for these higher-level factors, it is essential to investigate whether monkeys use the same higher-level factors that drive human eye-movement behavior.

Here we record viewing behavior of four monkeys freely viewing a set of images and compare eye-movement behavior to that of 106 human observers who freely viewed the same images in previous studies (Açik et al., 2010; Onat et al., 2014). We train salience models for each species and compare the patterns of feature influence between species. In addition we estimate high-level influences in view-

ing behavior by comparing fixation locations to locations marked as interesting by independent observers. We find that bottom-up salience influences are remarkably similar between both species. However, interestingness is not a good predictor for monkey eye-movements. Instead, image locations that are likely to be fixated by both species are better predicted by bottom-up saliency than by interestingness.

5.2 Methods

Participants

Eye movements were recorded from four rhesus monkeys (Macaca mulatta, three male). Recordings were carried out in accordance with National Institute of Health guidelines and were approved by the Emory University Institutional Animal Care and Use Committee. Eye movement recordings from humans come from two previous studies that used the same stimuli and comparable tasks. We analyzed data of 106 observers, 58 from Açik et al. (2010) and 48 from Onat et al. (2014).

Stimuli

Stimuli consisted of 192 images from three different categories (64 images in each category). "Natural" scenes were taken from the "McGill Calibrated Color Image Database" and depict mainly bushes, flowers and similar outdoor scenes. "Urban" scenes depicted urban and manmade scenes taken around Zürich, Switzerland. "Fractal" images were taken from Elena's Fractal Gallery, Maria's Fractal Explorer Gallery, and Chaotic N-Space Network available on the internet and depicted computer generated fractals. Figure 1 shows an example stimulus from the "Manmade and Urban scenes" category. Please see Açik et al. (2010) for more details.

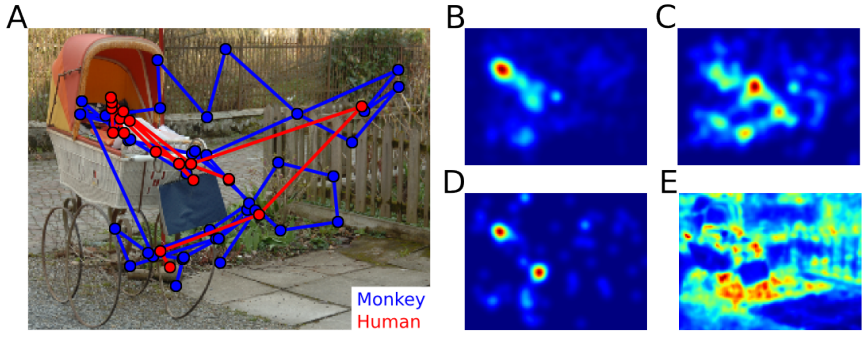


Figure 5.1: A) An example image with a monkey eye-trace (blue) and one human eye-trace overlaid. The four panels to the right show four different predictors for this image. B) A human fixation density map generated from all fixations except those in the example image. C) A monkey fixation density map generated from all fixations except those in the example image. D) Interestingess score generated by 32 independent raters. E) The prediction of a bottom-up salience model trained on human data.

Apparatus

Monkey eye movements were recorded with an ISCANN infrared eye-tracking system while each monkey sat in a dimly illuminated room. Monkeys were head posted during recordings. Stimuli were presented on a CRT Monitor with a resolution of 800x600 pixels and a refresh rate of 120Hz. The viewing distance was 60cm. Recordings were carried out at the Yerkes National Primate Center in Atlanta, GA, USA. Human eye movements were recorded with an EyeLink 1000 system (Açik et al., 2010) or an EyeLink II system (Onat et al., 2014). Human eye-movement recordings were carried out at the University of Osnabrück, Germany. Onat et al. (2014) presented stimuli on a CRT Monitor with a resolution of 1280x960 pixels and a refresh rate of 85Hz. The viewing distance was 80cm. Açik et al. (2010) used a 60Hz TFT screen with the same resolution and a viewing distance of 65cm.

Procedure and Task

Monkeys performed a free viewing task and were not explicitly rewarded for image viewing. Images were shown until a total looking time inside the image of 10s had accumulated. Between free viewing trials a color change task was carried out. It required the monkey to fixate a small rectangle (0.3°) that appeared at various locations on the screen. The rectangle changed color from gray to an equiluminant yellow at a randomly chosen time between 500 and 1,100ms. Upon release of its touch bar within 500ms after color change chow was delivered as reward (Jutras, Fries, & Buffalo, 2009; Jutras & Buffalo, 2010).

Recordings were carried out on three consecutive days. This kept sessions short enough for monkeys to attend to all images without loosing interest. On the first two days 66 randomly sampled images were shown twice and on the last day 60 images were shown twice. After 11 new images had been shown the same images repeated again. The order of presentation was the same for all monkeys. Due to a technical error, the data from one day from one monkey was discarded. To increase the amount of available data, and to potentially compare effects of memory later on, two monkeys repeated the experiment after four weeks.

Human observers were instructed to "freely view" the same images for six seconds (Onat et al., 2014). In contrast, Açik et al. (2010) presented images for five seconds. They used a template match task that asked observers to judge if an image patch was taken from the image presented just before.

5.3 Analysis

Data pre-processing

Saccade detection for humans was based on three measures: eye movement of at least 0.10°, with a velocity of at least 30°/s and an acceleration of at least 8000°/s2. After saccade onset, minimal

saccade velocity was 25°/s. Saccade detection for monkeys was carried out similarly but we additionally required that each saccade lasted at least 21ms and travels at least 0.35° of visual angle. This was necessary to cope with the lower sampling rate of the ISCANN system (240Hz vs. 500Hz). Samples in between two saccades were labeled as fixations.

To calibrate the monkey eye-tracking data we used the color change trials in between picture presentations. Since the color-change was subtle, monkeys had to fixate the rectangle until it changed color. We fitted a 2D affine transformation (least-squares fit) between average eye position after onset of the color change rectangle and the position of the rectangle in visual space. This took care of translations and skew in the monkey eye-tracking data. Human eye tracking data did not need to be re-calibrated as the eye-tracker was calibrated with a 9 point calibration grid before the experiment started.

Since the stimulus presentation time was different between experiments (5, 6 and 10s) we only used the first five seconds of image viewing for subsequent analysis. Additionally we rescaled the human eye-tracking data to the stimulus sized used for monkeys (800x600px).

Performance measure

This study investigates how well different factors predict fixation locations of humans and monkeys. Specifically we are interested in the predictive power of bottom-up-salience, within-species consistency, cross-species consistency and interestingness. These factors were quantified by "predictors" (described in detail below) that assign a score to every location in an image, which scales with the predicted likelihood of fixating this location. To evaluate if actual fixation locations received higher scores than non fixated control locations we computed the area under the receiver operating characteristic curve (AUC).

ROC curves for each predictor were generated by classifying actual fixations and control locations as fixated or non-fixated based on

the respective score at actual and control locations. To account for the center bias of fixations, control locations were drawn from the spatial bias of each observer (Tatler, 2007; Tatler et al., 2005, 2006). The classification was done with a simple threshold procedure. The ROC curve was generated by plotting, for all possible thresholds, the true positive rate of this classification versus the false alarm rate. The area under the curve is 1.0 if the classification is perfect, i.e. the distributions of score values at actual and control locations are perfectly separated. A value of 0.5 indicates a classification at chance level. Perfect misclassification results in an area under the curve of zero. Each predictor was evaluated for every observer and averaged over all stimuli within a category. This yielded one AUC per predictor, observer and stimulus category.

Feature-fixation correlations

To compare the influence of image features on viewing behavior we computed how well different features predicted fixation locations. In total we computed 16 different features on three spatial scales. First we computed luminance (LUM), blue-yellow (BY), red-green (RG) and saturation (SAT) channels of all stimuli. A first group of features represents a smoothed (Gaussian filter, $\sigma \in 2.1°, 4.2°$, 8.4°}) version of these simple features. A second group of features is computed by computing contrast in a Gaussian circular aperture ($\sigma \in \{2.1°, 4.2°, 8.4°\}$, suffix 'C' in Figure 2) on the luminance, blue-yellow, red-green and saturation channels. Contrast was computed according to the following formula:

$$C = \sqrt{(G(X^2, \sigma) - G(X, \sigma)^2)}$$

Where G convolves the input X with a Gaussian kernel with standard deviation σ. The third group represents the contrast of the contrast maps ($\sigma \in \{10.5°, 21°, 42°\}$, suffix 'TC' in Figure 2) defining the texture contrast. A fourth group consists of an edge detection filter (Sobel filter, 'SOBEL') and the intrinsic dimensionality 0-2 (ID0,

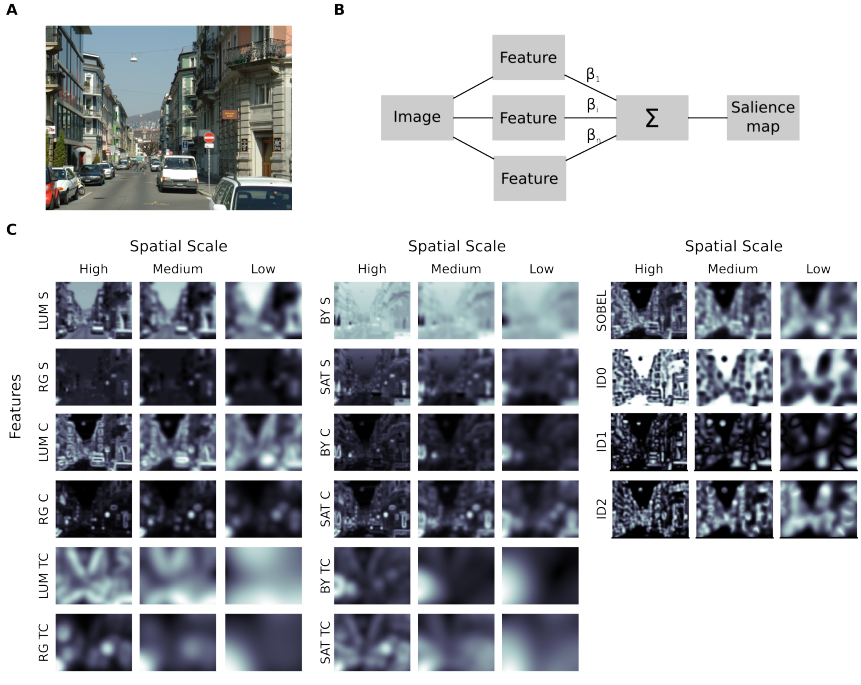


Figure 5.2: Bottom up salience model. A) Example image. B) Diagram of the bottom up saliency model. Weigths (βi) are determined by a logistic regression that predicts whether a location was fixated. C) Features for example image in A. S = smoothed, C = contrast, TC = texture contrast, LUM = luminance, RG = reed-green, BY = blue-yello, SAT = saturation, ID = intrinsic dimensionality.

ID1, ID2) of local image patches (c.f. Onat et al., 2014). Intrinsic dimensionality describes how well a local patch is described by an edge, corner or surface. Each feature was computed on three different spatial scales. This was achieved by down sampling of the original stimulus with a Gaussian pyramid up to two times. Each step of the pyramid halves the length and the width of the stimulus yielding the high (no down sampling), medium (once) and low (twice) spatial scale. Figure 2 shows all features for one example stimulus.

We then computed the AUC for predicting fixation locations with each feature. This yielded a vector of 48 AUC values for each observer and category. To assess the similarity in feature-fixation AUCs between monkeys and humans, we repeatedly (N=1000) regressed the pattern of average monkey feature-fixation AUCs onto an average of feature-fixation AUC vectors from four randomly sampled human observers. Sub-sampling AUC values from human observers allowed us to partially estimate the variance introduced by only having four monkey observers.

AUC values <0.5 indicate that a feature is anti-predictive of fixation locations, i.e. would be predictive if feature maps were multiplied with -1. This has an important implication for comparing features between species. If a feature is more predictive in one species, it's AUC is larger than 0.5 in both species and additionally larger in one of the two species. A feature that is more anti-predictive in one species has an AUC smaller than 0.5 in both species and a smaller AUC for the species for which it is more anti-predictive. In a scatter plot matching the AUCs of species against each other, such features are located somewhere off the diagonal. If features are systematically more predictive or anti-predictive in one species, AUCs fall on a line that pivots around 0.5. Regression coefficients larger than one therefore indicate that predictive features are more predictive and anti-predictive features are more anti-predictive. If the linear regression additionally explains most of the variance in feature-fixation AUCs, a linear relationship between feature influences exists be-

tween species. We thus tested whether regression coefficients were different from one with a two-sided t-test and computed the average variance explained by the regression models.

Salience Model

The details of the salience model have been described previously (Wilming et al., 2013). Briefly: The bottom-up salience model consisted of a weighted linear combination of the set of 48 features described above. Figure 2B shows a schematic drawing of the salience model. Weights were obtained by a logistic regression that predicted whether an observation (vector of feature values) was taken from a fixation or from a control location. Control locations were sampled from the spatial bias of fixations. Weights were obtained for each species and stimulus category separately. To evaluate the performance of the saliency model we performed a leave-one-out cross validation on the level of observers. This ensured that data from the observer to be predicted was never used during weight estimation. Thus weights were estimated with data from other subjects, the procedure thereby focused on singling out bottom-up influences that are shared by observers in one species.

Within- and cross-species consistency

How well fixation locations from one species predicted locations from the other was quantified with a cross-species predictor. The cross-species prediction for a specific image was generated by smoothing all fixations from one species on that image with a Gaussian filter of FWHM=2° and subsequently normalizing the 2D map to unit volume (Wilming, Betz, Kietzmann, & König, 2011). This yielded a score for each location in a visual stimulus that was used to compute the AUC for predicting fixation locations of the other species.

The within species prediction was similar but used fixations from the own species (without the subject currently evaluated). This forms an effective upper bound for the predictability of viewing

behavior (Wilming, Betz, Kietzmann, & König, 2011).

Interestingness

Onat et al. (2014) investigated whether viewing behavior in humans is driven by bottom-up salience or higher level factors that are independent of salience. To this end, they asked 32 observers to select five "interesting" locations on each of the stimuli used here. Observers selected five locations with a pointing device in a self-paced manner. We used these data to define an "interestingness" predictor in close analogy to Onat et al. That is, we smoothed all interesting locations with a Gaussian filter of FWHM=2° and normalized the 2D map to unit volume. Figure 1 shows examples of the four predictors for one stimulus.

Comparison of cross-species prediction with interestingness and bottom-up salience

We were interested whether locations that were fixated by both species were correlated with high bottom-up salience values or high interestingness values. To this end we interpreted the prediction of each species for a specific image as the probability that each location was fixated. This allowed us to compute the probability that each location would be fixated by both species ($P(Fix|Location, Human) * P(Fix|Location, Monkey)$). In a next step we compared this probability map to the interestingness prediction and the bottom-up salience prediction for the same image. Since both predictions have a different scale we normalized each to unit area. We then weighted both predictions by the probability that a location was fixated by both species. Thus, locations that were not fixated by both species were suppressed in both predictions. Finally we asked whether weighted interestingness or weighted bottom-up salience had higher values by entering both predictions into an AUC analysis. We thereby compared, on each stimulus, the distributions of both predictors at locations that were fixated by both species. This yielded one AUC

value per stimulus and category. In analogy to the other AUC values, we used salience values as actuals and interestingness values as controls. AUC values larger than 0.5 imply that salience has higher values than interestingness at locations that were fixated by both species. AUC values smaller than 0.5 imply the opposite.

Statistical comparisons

To address the different number of observers for each species we bootstrapped 95% confidence intervals (CI) for human AUC values by repeatedly selecting four random human observers and averaging their AUC values. This estimated the distribution of average AUC values to be expected had only four human observers participated in the experiment. Mean monkey AUC values falling outside of the 95% CI for humans were interpreted as a significant difference.

For human observers, within species comparisons were carried out by a bootstrapping procedure that estimated the sampling distribution of AUC differences. We subtracted the AUC values in question for all observers. We then repeatedly (N=5000) computed the average of 106 (one for each human observer) randomly sampled (with replacement) AUC differences. We deemed a comparison to be significant when the 95% CI of sampled differences did not include 0 (Bonfferoni-Holm corrected per stimulus category).

Previous studies have shown that estimates for within-species consistency AUC values are dependent on the number of observers used for prediction (Wilming, Betz, Kietzmann, & König, 2011). Fewer observers produce smaller estimates than larger group sizes. When comparing humans to monkeys any difference in within-species consistency is therefore potentially due to different sample sizes.

To account for this possibility we estimated the effect of smaller group sizes in our human data. We sampled groups of observers with $N =\in 4, 16, 32$ observers 500 times (2000 groups). In each group we estimated the consistency between human observers. We were also interested in the development of the performance of within-humans consistency over the number of observers used for com-

puting the prediction. We therefore additionally subsampled each group and predicted each observer in the group with different numbers (1,2,3,4 or 16) of observers from the same group. For example, if the group size was 16, we predicted each observer with 1,2,3, and 4 randomly sampled observers. This yielded 500 estimates for each group size and number of predicting observers which we used to determine intervals that contained 95% of the samples around the estimated mean. Since the estimated mean appeared to follow an exponential function across the number of observers used for prediction we also fitted an exponential function to the data for each group size. This was only done to ease visualization of the data.

5.4 Results

We started by investigating the influence of bottom-up salience in humans and monkeys. To this end we compared how well different image features predicted fixation locations from both species by comparing feature values at fixated and control locations (Fig. 3). Panels A and B show eye-movement trajectories on one stimulus and one example feature map. Panel C shows the distribution of feature values at actual fixation locations and control locations for one exemplary feature (ID2). The area under the receiving operating characteristics curve for these two distributions gives a score of predictability (AUC). Panel D shows AUC values for all features on urban scenes for both species.

The average deviation of AUC values from 0.5 across image features ($\langle |AUC_{ftr} - 0.5| \rangle_{ftr}$) for urbans and fractals is significantly lower in humans ($\langle AUC \rangle_{urbans} = 0.06$, $\langle AUC \rangle_{fractals} = 0.05$) than in monkeys ($\langle AUC \rangle_{urbans} = 0.07$, $\langle AUC \rangle_{fractals} = 0.06$, $p_{urbans} = 0.006$, $p_{fractals} = 0.001$, paired t-test). Monkeys' fixations were therefore better predicted by image features. We furthermore wanted to know, whether those features that predicted humans well also predicted monkeys well (Figure 3D). We averaged feature-fixation AUC values over monkeys and used the resulting AUC values to

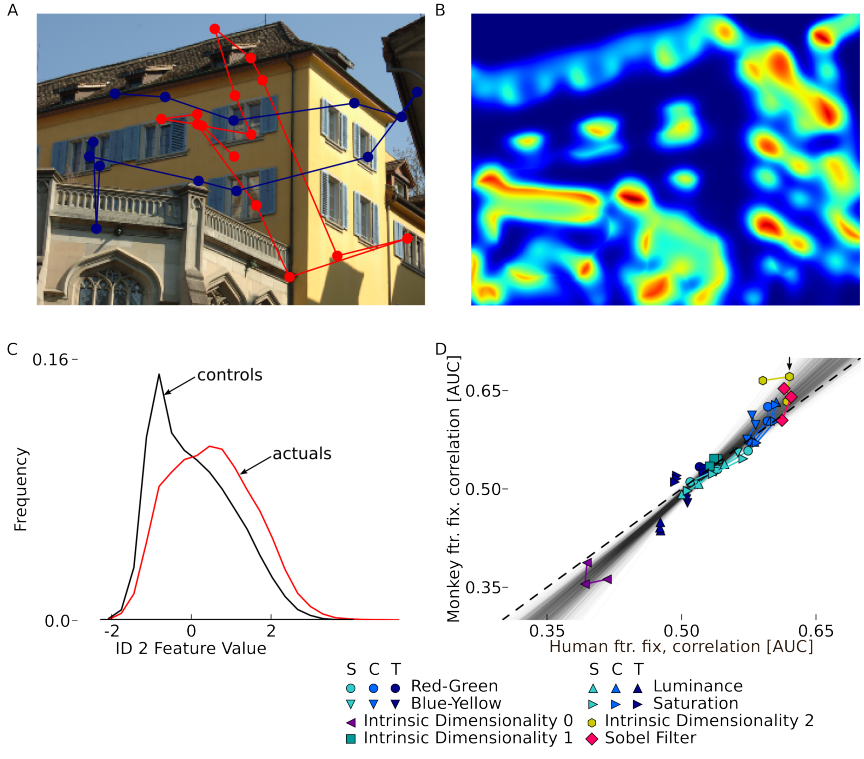


Figure 5.3: A comparison of Feature-fixation AUC values for urbans scenes. A) Example image with eye movement traces from one human (blue) and one monkey (red) observer. B) intrinsic dimensionality 2 feature map of the example image. C) Histograms of intrinsic dimensionality 2 feature values at control locations (black) and human fixation locations (red). ROC curves are generated by computing true positive and false positive rates at every possible threshold. D) Feature fixation AUC values for humans and monkeys for different features. Symbols and colors show different image features. Different spatial scales of a feature are indicated by connecting lines between data points. Transparent black lines show linear regression fits between both species.

predict average feature-fixation AUC values from four randomly sampled human observers. The monkey feature-fixation AUC values explained 91% and 94% of the variance of human feature-fixation AUC values on urban scenes and fractals respectively. Hence, the patterns of feature-fixation AUCs were very similar between monkeys and humans on urban scenes and fractal scenes. A linear regression between feature-fixation AUC values (Figure 3D) showed a slope significantly larger than 1 for urban scenes and fractals ($\langle\beta_{urbans}\rangle = 1.23$, $p(\beta_{urbans} = 1)<0.0001$; $\langle\beta_{fractals}\rangle = 1.43$, $p(\beta_{fractals} = 1)<0.0001$). This gives further evidence that the pattern of effective image features is highly similar in the two species, i.e. the same features predict eye-movements, and that it is slightly more effective in predicting monkey fixation locations.

As a next step we analyzed the performance of a complete bottom-up salience model. Black bars in Figure 4 show the mean performance of the bottom-up salience model for urban and fractal scenes (urbans: $\langle AUC\rangle_{humans} = 0.67$, $\langle AUC\rangle_{monkeys} = 0.69$; fractals: $\langle AUC\rangle_{humans} = 0.65$, $\langle AUC\rangle_{monkeys} = 0.67$). Relative to the best feature, the salience model for humans gained performance through the combination of different features, but the best feature for monkeys was almost as good as the final salience model (humans: $\langle salience-sobel\rangle_{urbans} = 0.05$, $p<0.0001$, $\langle salience-sobel\rangle_{fractals} = 0.023$, $p<0.0001$, paired t-test; monkeys: $\langle salience - ID2\rangle_{urbans} = 0.002$, n.s., $\langle salience - ID2\rangle_{urbans} = 0.002$, n.s.). This is consistent with the observation that the average bottom-up salience AUC value for monkeys was higher but was still contained within the 95% confidence interval for humans (Figure 4, black bars). Bottom-up salience therefore showed comparable performance for human and monkey observers on this dataset.

In addition to bottom-up salience we considered interestingness, i.e. locations labeled as interesting by independent human observers. Interestingness was the best predictor of human eye-movement behavior ($\langle AUC\rangle_{urbans} = 0.75$, $\langle AUC\rangle_{fractals=}0.70$), almost reaching the average inter-observer consistency between human observers

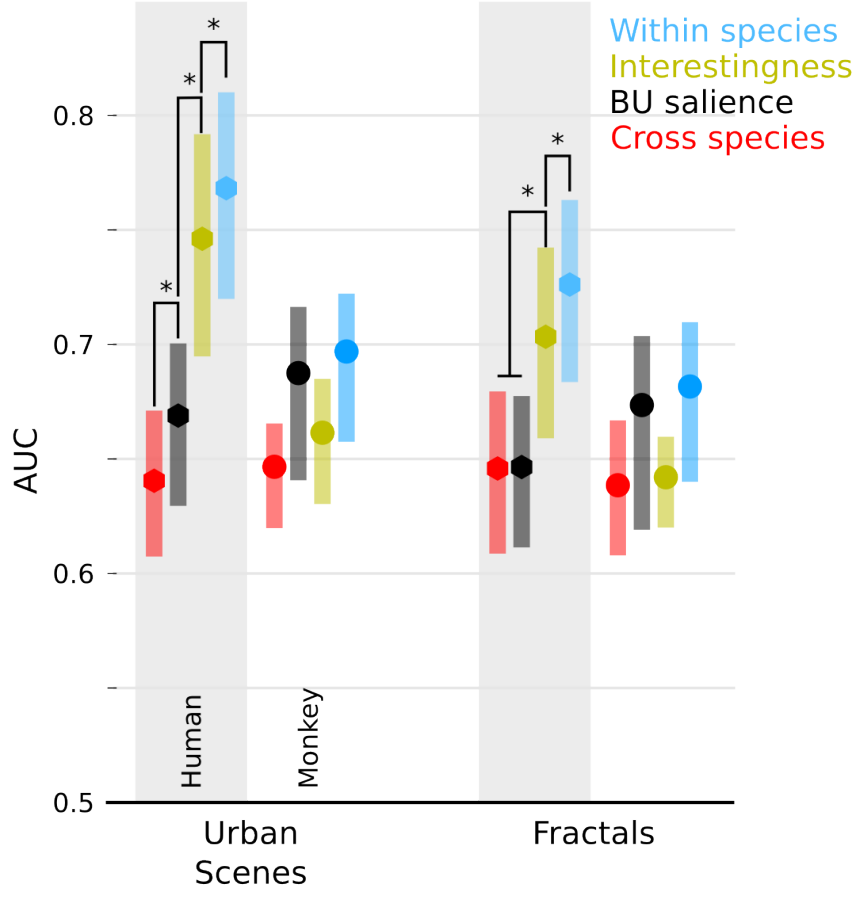


Figure 5.4: Predicting monkey and human fixation locations with different predictors. Each bar shows the mean value and 95% CIs for one predictor. Confidence intervals are computed by repeatedly sampling four observers to allow better comparison between human and monkey data. Values for predicting human observers are shaded in gray. Comparison between predictors for humans are within-observer comparisons of 106 observers. For urban scenes, all pairwise comparisons are significant with $p<0.0001$. For fractal scenes, the difference between bottom-up salience and cross-species is not significant. All other comparisons are significant with $p<0.0001$.

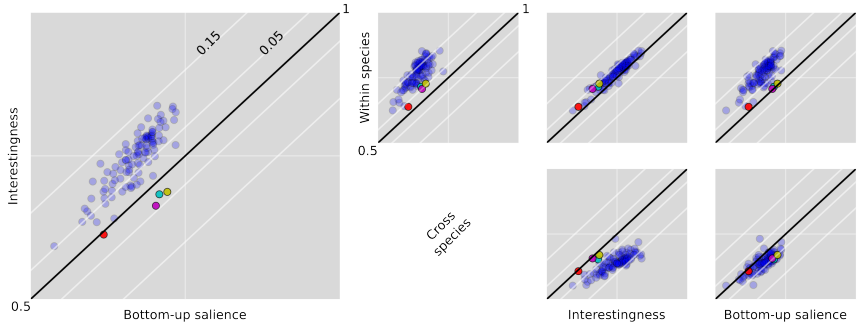


Figure 5.5: AUC values for predicting fixation locations of individual observers on urban scenes (same data as in Figure 4). Each dot shows the AUC value for predicting one observer. Humans are shown in blue, individual monkeys are represented by other colors. Each panel compares how well two different factors predict the same individuals. Faint white diagonal lines show AUC differences of 0.05 and 0.15.

($\langle AUC \rangle_{urbans} = 0.77$, $\langle AUC \rangle_{fractals} = 0.72$, difference to interestingness is significant in both cases, $p<0.001$, bootstrapping procedure).

Interestingness predicts human viewing behavior much better than bottom-up salience ($\langle interestingness - salience \rangle_{urbans} = 0.1$, $\langle interestingness - salience \rangle_{fractals} = 0.08$, $p\langle 0.001$ in both cases). This suggests that interestingness predicts those fixation locations that are selected because of higher-level top-down control. In turn, the high predictive power of interestingness suggests that these higher-level factors play an important role in the selection of gaze locations in humans (c.f. Onat et al., 2014).

Turning to the predictive power of interestingness for monkey observers, we found that interestingness was, on average, a worse predictor for monkey fixation locations than bottom-up salience ($\langle AUC \rangle_{urbans} = 0.66$, $\langle AUC \rangle_{fractals} = 0.64$). Interestingness had lower AUC values for all four monkeys on urban images and for three out of four on fractals (Figure 5). Locations that are labelled as interesting by humans were therefore fixated often by humans but not by monkeys. The fact that bottom-up salience predicted both

species to a comparable extend, but interestingness did not, suggests that both species were driven by the same bottom-up salience system but not by the same higher-level factors.

This leads to the prediction that if human and monkey observers share the same bottom-up salience system, each species should target, to some extend, locations that are highly bottom-up salient. Indeed, we found above chance cross-species prediction performance that was comparable across category and predicting species ($\langle$ human $\rightarrow$ monkey $\rangle_{urbans}$=0.64, $\langle$ monkey $\rightarrow$ human $\rangle_{urbans}$=0.65, $\langle$ human $\rightarrow$ monkey $\rangle_{fractals}$=0.64, $\langle$ monkey $\rightarrow$ human $\rangle_{fractals}$=0.64, n.s., bootstrapping procedure).

We furthermore observed that the bottom-up salience model is more, or equally predictive as the cross-species prediction for urbans ($\langle salience - across \rangle_{humans} = 0.03$, $p < 0.001$; $\langle salience - across \rangle_{monkey} = 0.04$, positive in four out four monkeys, Figure 5) and fractals ($\langle salience - across \rangle_{humans} = 0.00$, n.s; $\langle salience - across \rangle_{monkey} = 0.04$, positive in three out four monkeys, Figure 5). Humans and monkeys therefore did not necessarily pick the same salient locations on each image. Human observers, for example, might have chosen salient locations that are also interesting, while monkeys picked different salient locations either by chance or because of unknown higher-level influences. But since interestingness explained viewing behavior in humans much better than bottom-up salience, it seemed plausible that the cross-species prediction mainly explained those fixations that were driven by bottom-up salience in both species.

To corroborate this hypothesis we compared interestingness predictions and bottom-up salience predictions in areas that were likely to be fixated by both species. We found that these locations had higher bottom-up salience values than interestingness values ($\langle AUC \rangle_{urbans} = 0.64$, $p < 0.001$, t-test). In sum, these findings suggest that the found similarities in viewing behavior between humans and monkeys can be better explained by bottom-up salience than by a set of shared top-down influences.

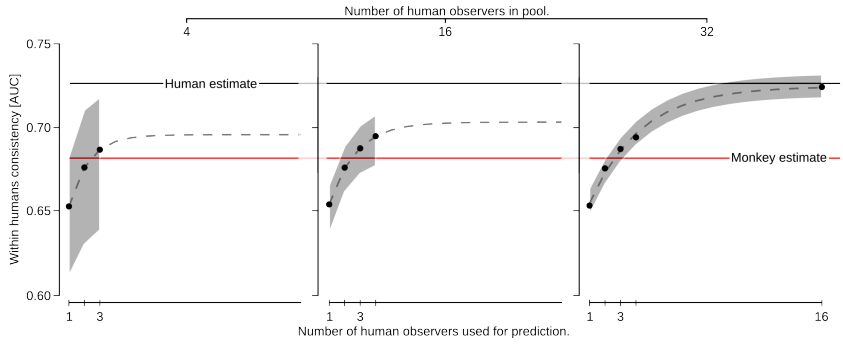


Figure 5.6: Dependency of within-humans consistency for fractal images on the number of observers used for prediction and number of observers available. Black dots show within-humans consistency estimates, dashed lines show best fitting exponential functions. Shaded areas depict confidence intervals that are generated by fitting exponential functions through confidence intervals for each data point. Panel on the left shows consistency estimates when predicting observers are drawn from a pool of size four. Center and left panel show results for pool sizes of 16 and 32 observers.

In principle it would be interesting to estimate how much of the consistent viewing behavior within a species can be explained by bottom-up salience and interestingness. The amount of consistent viewing behavior is usually (Wilming, Betz, Kietzmann, & König, 2011) estimated by computing the inter-observer consistency (here within-species consistency). We found that humans had higher within-species consistency than monkeys ($\langle humans \rangle_{urbans} = 0.77$, $\langle humans \rangle_{fractals} = 0.73$, $\langle monkey \rangle_{urbans} = 0.70$, $\langle monkey \rangle_{fractals} = 0.68$). We found especially intriguing that, for monkeys, bottom-up salience reached 95% of the within-monkey consistency. This suggested that bottom-up salience explains most of the consistent viewing behavior between monkeys.

However, Wilming, Betz, Kietzmann, and König (2011) have shown that the within-species consistency depends on the number of observers that are used for estimating the prediction. Specifically, they observed that fewer observers lead to smaller consistency values.

This implies that the within-monkey consistency might be larger if data from more monkeys were available. To allow a more meaningful interpretation of the within-species consistency, we sub-sampled the human observers into smaller groups. Figure 6 shows the results for groups of size 4, 16 and 32 observers for the viewing of fractal images. As expected, the consistency between human observers increased with the number of observers used for the prediction. Further, the smaller the group of observers, the larger is the variance of the estimate. To compare the within-human and monkey consistency we compared the case where three human observers predicted one remaining observer and where all observers came from a pool of four observers (left panel, third dot). This case is in close analogy to the within-monkey estimate. The monkey estimate was well within the confidence interval for the human estimate. However, since the confidence intervals for the human estimate covered a large range (0.64-0.72) this mainly indicated a lack of power to distinguish credible differences in within-species consistency. Whether or not the within-species consistency is truly different between the species, and whether bottom-up salience explains most of the consistent viewing behavior, can therefore not conclusively be decided with the present data.

We now turn to the viewing of natural scenes from our dataset. In principle AUC values on natural scenes showed the same pattern as on urban and fractal scenes. For humans, interestingness ($\langle AUC \rangle = 0.65$) was close to the within-species prediction ($\langle AUC \rangle = 0.66$), followed by bottom-up salience ($\langle AUC \rangle = 0.59$) and the cross-species prediction ($\langle AUC \rangle = 0.56$). Monkeys showed the same order, except that interestingness was the worst predictor for eye-movement behavior ($\langle AUC \rangle_{interestingness} = 0.58$, $\langle AUC \rangle_{within} = 0.62$, $\langle AUC \rangle_{bottom-up} = 0.58$, $\langle AUC \rangle_{cross-species} = 0.57$). However, we found that all predictor scores were significantly lower compared to the other scene types. This is consistent with Wilming, Betz, Kietzmann, and König (2011) who report that natural scenes are most influenced by the central fixation bias and show much less

inter-observer agreement than urban scenes.

5.5 Discussion

The present study reveals remarkably similar influences of bottom-up salience on oculomotor behavior in humans and macaque monkeys in a free viewing paradigm. With our analysis of feature-fixation patterns we probed the influence of different visual features in guiding eye-movement behavior. These patterns were almost identical between monkeys and humans on categories that elicit consistent and stimulus dependent viewing behavior.

Previously Einhäuser, Kruse, et al., 2006 reported that bottom-up salience has a small influence ($\langle AUC \rangle = 0.59$) in monkeys during free viewing of natural scenes. This is consistent with monkey behavior that we observed on natural scenes. But, we also demonstrated that bottom-up salience was more effective for predicting monkey fixation locations on urban scenes and fractals. Bottom-up salience is therefore likely to be more prominent in guiding monkey eye-movements than previously thought.

We find that the predictive power of bottom-up salience was slightly higher in monkeys than in humans. This difference between species was, however, not significant. A small difference between species is in line with results reported by Einhäuser, Kruse, et al. (2006) and Berg et al. (2009). However, in our study monkeys were better predicted than humans, while Berg et al. (2009) report that humans are better predicted than monkeys. In principle it is entirely possible that the difference in the influence of salience is due to the different stimulus types (still images vs. videos respectively) used. We would like to point out though, that the small number of monkeys (4, 4 and 5) in all three studies makes conclusive inference about the small effect sizes observed very difficult. In any case, we found that differences between stimulus categories were much larger than differences between species. If there are genuine differences in bottom-up salience, their behavioral relevance are likely mitigated

in comparison to between species differences in other factors.

Whether or not bottom-up salience has a causal influence on eye-movements is a matter of ongoing debate (Einhäuser & König, 2003; Henderson, Brockmole, Castelhano, & Mack, 2007; Einhäuser et al., 2008; Schütz, Trommershäuser, & Gegenfurtner, 2012). The alternative is that bottom-up salience covaries with unknown higher-level factors that causally drive eye-movements. And indeed, it would be naïve to assume that our bottom-up salience model exactly mimics salience computations in the brain. But our findings limit the potential factors that cause viewing behavior and covary with bottom-up salience. Because bottom-up salience explains similar amounts of viewing behavior in both species, and feature-fixation AUC patterns were similar in both species, bottom-up salience would most likely covary with factors that are shared between species. This makes it unlikely that human semantic scene interpretation, for example inferred intentions and emotions, is driving feature-fixation AUCs. More likely factors might be bottom-up salience or the detection of objects (Einhäuser et al., 2008; Nuthmann & Henderson, 2010).

Where could be the origins of bottom-up-salience similarity in humans and monkeys? The location of salience computation has neither been conclusively pin-pointed in monkeys or humans. Putative answers range from sub-cortical areas like the pulvinar to areas in early visual cortex (V1 Li & Zhaoping, 2005; V4 David, Hayden, Mazer, & Gallant, 2008) and higher level areas like LIP and FEF (Mazer & Gallant, 2003; Li & Zhaoping, 2005; Arcizet et al., 2011b; Bogler et al., 2011). For many higher-level areas it is debated whether exact functional homologue areas exist in monkeys and humans. In early visual cortex, at least areas V1 and V2 are thought to be closely related in humans and monkeys (Mantini, Corbetta, Romani, Orban, & Vanduffel, 2012; Orban, Van Essen, & Vanduffel, 2004). We think it is reasonable to assume that the homology of early visual cortex contributes to the similar feature-fixation AUCs observed in this study. Alternatively, the superior colliculus could play a decisive role (Krau-

zlis et al., 2013; Gattass & Desimone, 2014; Zénon & Krauzlis, 2012). The homology of early areas and comparable feature-fixation AUCs are consistent with the assumption that features for a bottom-up saliency map are computed in early visual cortex.

We found differences in the predictive power of bottom-up salience and interestingness between species. Bottom-up salience explained monkey viewing behavior better than interestingness, while interestingness explained human fixation locations much better than bottom-up salience. Given comparable bottom-up salience AUCs between species these results indicate that humans share a set of guiding factors, as indexed by interestingness, that are missing in monkeys.

The interestingness scores come from independent human observers who are likely guided by bottom-up salience, oculomotor biases and higher-order factors (e.g. semantic scene interpretation) when they mark locations as interesting. Interestingness can therefore be interpreted as a combination of bottom-up salience influences and additional higher-level factors. Since the bottom-up system appears to be similar in both species, the difference in the predictive power of interestingness can be attributed to additional factors that guide the selection of interesting points. Furthermore the use of control locations from the spatial bias of observers effectively controls for the spatial oculomotor bias. The observed difference in this study therefore suggests that humans are strongly guided by factors that are independent of bottom-up salience and oculomotor biases. The low predictive power of interestingness for monkeys suggests that these higher-level factors do not guide fixation selection in monkeys.

One might argue that the stimuli used in this study were not ecologically valid for monkeys (c.f. Einhäuser, Kruse, et al., 2006). A monkey defined interesting score might have predicted monkeys much better than humans. Viewing of faces and social scenes, for example, triggers specific viewing behavior across monkeys (McFarland et al., 2013; Guo, Robertson, Mahmoodi, Tadmor, & Young, 2003, 2009). While this is clearly plausible, the predictive power of

interestingness and bottom-up salience were also reversed between species for fractals. A scene category that is supposedly not ecologically valid for either species. Thus at least some ecologically invalid stimuli trigger higher-level factors that guide viewing behavior in humans but not in monkeys.

Unfortunately the small number of monkeys prohibited us from reliably estimating the within-monkey consistency. At face value it matched the predictive power of the bottom-up salience model, suggesting that bottom-up salience completely explains consistent viewing behavior across monkeys. If this was the case we could have concluded that no additional higher-level factors are necessary to explain consistent viewing behavior. However, the number of monkey subjects is small. This is a direct consequence of the tremendous effort involved in physiological experiments involving monkeys. More often than not such studies involve just two subjects. Hence the number of monkeys available for concurrent psychophysical experiments is limited as well. The within-monkeys consistency is potentially underestimated. This leaves open the possibility that monkeys are driven by consistent higher-level factors, which are different to human higher-level factors. As a remedy, joint multicenter studies involving identical paradigms and stimuli should be conducted.

We started out to investigate if macaque monkeys rightfully are the primary model system for studying attention and oculomotor control. The similarity of feature-fixation AUCs suggests that monkeys are an excellent model system for studying the generation and functioning of bottom-up salience. We suggest that humans and monkeys use a similar bottom-up salience that potentially corresponds to homologue areas in early visual cortex.

We found that human observers were guided by salience independent factors that were not present in monkeys. Whether or not monkeys are a good model system to investigate how higher-level factors guide fixation selection depends on how they emerge in humans. It might be that humans and monkeys use the same ma-

chinery to guide eye-movements, but that humans simply "have more of it". For example, if interestingness is interpreted as reflecting the influence of higher level semantic features, such as object related information, it is plausible that the inferior temporal cortex contributes objet specific information (Hung, Kreiman, Poggio, & DiCarlo, 2005). Indeed, this structure is larger in humans (Passingham, 2009). Nevertheless Kriegeskorte, Mur, Ruff, and Kiani (2008) report that monkey and human IT share a very similar categorical representation. This suggests that monkeys possess similar capabilities to categorize scenes as humans. It thus seems plausible that the same neuronal mechanisms are at work in monkeys and humans. At the same time, fewer and less clear homologous brain structures exist after early visual cortex (Orban et al., 2004). Human LIP for example possesses more retinotopically organized areas than monkey LIP (Patel et al., 2010). And on a larger scale oculomotor control in the human brain appears to be more lateralized (Kagan, Iyer, Lindner, & Andersen, 2010; Oleksiak, Postma, van der Ham, Klink, & van Wezel, 2011) while the monkey brain shows more contralateral specificity (Kagan et al., 2010). It is therefore equally plausible that the human brain possesses genuinely different mechanisms to drive eye-movement behavior. Whether or not findings about higher-level guidance of eye-movements in monkeys generalize to humans is therefore presently unclear.

To sum up, we found that both species are equally guided by bottom-up salience and that monkeys are most likely an excellent model system for the study of salience. Crucially humans show additional consistent viewing behavior between observers, which could not be observed in monkeys. One of the key questions that remains, is if the same neuronal mechanisms produce different behavior in monkeys and humans.

Chapter 6

Predictions in the light of your own action repertoire as a general computational principle

The chapter has been published in Behavioral and Brain Sciences:
König, P., Wilming, N., Kaspar, K., Nagel, S. K., & Onat, S. (2013). Predictions in the light of your own action repertoire as a general computational principle. *Behavioral and Brain Sciences*, *36*(03), 39–40
A comment to:
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204

Present Cognitive Science is characterized by a dichotomy separating sensory and motor domains. This results in a perceived gap between perception and action and is mirrored in leading theories of cognition. For illustration we consider the visual neurosciences, a paradigmatic field for the investigation of sensory processes. A discourse given by standard textbooks depicts a world external to the agent with a set of pre-given attributes and objects. Sensory processing starts with transmitting these attributes by low-level neurons to subsequent stages. There more elaborated computations extract patterns of stimulus features and objects. Up to this point, processing focuses on a veridical representation of the external world, serving for later decisions and actions. We argue in favor of a radical change of this view, assigning a central role to predictions of sensory consequences of ones own actions and thereby eliminating the strict separation of sensory and motor processing.

In the target article Andy Clark beautifully describes the central role of predictions in sensory processing. We endorse this view; yet two complementary aspects are needed. Firstly, predictability of sensory signals serves as a normative principle guiding sensory processing and as a boundary constraint in the selection of information to process. Secondly, predictions are performed only in the context of the own action repertoire (König & Krüger, 2006). These two specifications have crucial implications.

The information content of the primary sensory signal is enormous and extraction of information without further constraint is an ill posed problem. However, it is not the task of the sensory systems to process all possible details, and a reduction of information is paramount. Even in simple model systems, taking into account a limited behavioral repertoire converts demanding sensory processing into a tractable problem (Wyss, König, & Verschure, 2004). Applying the normative principle of predictability generalizes this idea and serves as a selection criterion for features to process and variability to ignore. Indeed, within the hierarchy of the visual system neuronal response properties are invariant to more and more parametric

changes of the sensory input (Tanaka, 1996). Even category learning at higher levels of the visual system can be interpreted within this framework. The commonalities between different instances of the same category relate to similar sensory-motor patterns generated by the interaction with these "objects". Finally, actions are directly related to the agent's survival and thereby processing features that change predictably given chosen actions are more relevant than those that do not. Hence processing of sensory signals is guided by the relevance for behavior and relevance is expressed by the ability to predict sensory changes contingent on the own action repertoire.

A paradigm that is based on the active interpretation of incoming sensory information such that it makes sense for the agent, intends to replace a passive representationalist view. In such a paradigm the predicted future state of the world is important in as far as it interacts with own actions and variables of importance are co-determined by the action repertoire. A demonstration of the integration of new sensory information (magnetic north) that is co-determined by own movements (yaw-turns) is given by the feelSpace project (Nagel, Carl, Kringe, Märtin, & König, 2005; Kärcher, Fenzlaff, Hartmann, Nagel, & König, 2012). Comparing different species, e.g. cat and human with similar visual input (Betsch, Einhäuser, Körding, & König, 2004; Einhäuser et al., 2009), the remarkable differences in the sensory hierarchy appear to be at odds with a passive representationalist view and await an explanation. Here, differences in behavioral repertoire offer themselves. Pointedly, we speculate that the huge action repertoire of humans, due for example to opposable thumbs, might foster the illusion of a veridical perception of the world. It has been emphasized early on, that cognitive and motor develop in parallel and mutual dependence (Piaget, 1953). To grow up means to harden specific action routines on the one hand, but to loose a bulk of alternative action capabilities and cognitive flexibility on the other hand. Furthermore, a large variability of perceptual interpretation of identical physical stimuli is found between humans of the same culture area as well as between different cultures (Segall, Campbell,

& Herskovits, 1963). A critical view at our own culture reveals many aspects that serve to increase the reliability of predictions. In summary, agents with identical sensory organs but different action repertoires might have a very different view of the world.

Is the concept of normative principles plausible in view of our knowledge of cortical networks? Neuronal computations are constrained by properties of the brain in the form of number of neurons and synapses, space and energy consumption. The latter has served as an argument for sparse coding, i.e. low mean activity at constant variance of activity (Barlow, 1961). The insight that receptive fields of simple cells in primary visual cortex form such an optimally sparse representation of natural images drastically increased interest in normative models (Olshausen & Field, 1996; Simoncelli & Olshausen, 2001). Properties of the second major neuron type in primary visual cortex, complex cells, can be understood along similar lines as optimizing stable representations (Körding, Kayser, Einhäuser, & König, 2004; Berkes & Wiskott, 2005). Importantly, both optimization principles can be easily implemented by recurrent connectivity within a cortical area (Einhäuser, Kayser, König, & Körding, 2002a). Hence, existing normative models of the early visual system are plausible in view of anatomical and physiological data.

A critical test of the concept will be the application well beyond processing in primary visual cortex. The step from sparseness and stability to predictability as an optimization principle requires critical extensions. Phillips, Kay, and Smyth, 1995 put forward a very promising proposal: Coherent infomax selects and coordinates activities as a function of their predictive relationships and current relevance. The relation of this approach (Phillips, 2012, this issue) to the free energy principle (Friston, 2010) and optimal predictability (König & Krüger, 2006) has to be investigated. These developments hold the promise to apply to "higher" cognitive functions as well and giving rise to a true theory of cognitive science.

Chapter 7

Discussion

In this thesis I have presented five manuscripts that focus on how humans decide where to look next. The individual chapters (except the last one) already contain a discussion of their results with respect to the fixation selection process. The thesis as a whole focuses on the saliency map model of attention. I hence focus here on a slightly different aspect: What have the manuscripts in this thesis contributed to our understanding of the saliency map model as a model for free viewing of pictures?

The introduction of this thesis and the contained manuscripts should have made clear that we have not (yet!) understood the fixation selection process to a satisfactory level. In what follows, I will relate results from this thesis to some ways in which our understanding is not satisfactory, and point out how we can advance our understanding of the fixation selection mechanism.

In chapter 2 we have shown that evaluating saliency models is a difficult enterprise, because the results are critically dependent on the dataset being used. In a very positive development, one recent benchmark (Judd et al., 2012) is now reporting lower and upper

bounds[1], as suggested in chapter 2. This leads to the question of whether existing saliency map models have already reached the upper bound.

Figure 1.4 in the introduction compares different models to the reference frame that we suggested in chapter 2. The best models reach about 75% of the inter-observer consistency (upper bound), thereby missing it by a fairly wide margin. At the same time the predictive power of the spatial bias of fixations (lower bound) reaches 68%, which makes it difficult to judge whether models are successful because they are a good model of stimulus-dependent saliency or of stimulus-independent spatial bias. Current saliency map models of attention are therefore not satisfactory explanations of free viewing behavior.

This is, of course, a pessimistic view of what has been achieved by the scientific community. A more constructive perspective can be had by asking what we have learned from the best models that predict fixation locations. For example, returning to Figure 1.4, why are some models better than others? Comparing three top-scoring models (Harel et al., 2007; Judd et al., 2009; Vig, Dorr, & Cox, 2014), we can see that these models reach high predictive power in different ways. Harel et al. (2007) suggest that their "Graph based visual salience" rates locations in the center of the image as more salient. Their model thereby matches the center bias of human fixations. Judd et al. (2009) incorporate the location of the horizon, faces and persons into their prediction. Vig et al. (2014) focus on identifying effective visual features from a large family of feature detectors. A commonality between the three models is that all include the center bias of fixations explicitly or implicitly. These observations allow two different conclusions. First, high predictive power can be produced by a variety of different computations of saliency. Second, incorporating the center bias into the prediction appears to be beneficial. Both these conclusions pose problems for using models as explanations for viewing behavior.

[1]Without our influence, the authors arrived at similar conclusions independently.

The latter conclusion resonates with previous findings (c.f. chapter 2, Tatler and Vincent, 2008; Tatler, 2007; Tatler and Vincent, 2009) that the central bias of fixations has predictive power for fixation locations. This is unproblematic when our goal is mere prediction. But when our goal is to understand the role of saliency in the process of selecting fixation locations, then the spatial bias is a strong confounding factor. A model that only exploits the spatial bias has no explanatory power beyond the spatial bias and therefore no added value for explaining viewing behavior. How do we decide which of the models is the better explanation of viewing behavior over and above the spatial bias? Our results suggest one way: We could remove the strong influence of the spatial bias by retesting the models on a data set with a very small lower bound and a large upper bound, such that the spatial bias has no predictive power. This would dissociate the spatial bias from other factors that guide viewing behavior.

The conclusion that different mechanisms for computing saliency can all yield high predictive power poses a bigger problem. How would we decide between different models that all still have similar predictive power after we have taken the spatial bias into account? The observation that high predictive power can be reached in different ways highlights that predictive power alone is not sufficient to reach veridical explanations via computational modeling. In chapter 2 we have suggested that the predictive performance of different models is a proxy for their explanatory power, but cautioned that models must also be plausible. We argued that to be a good explanation a model must be predictive and plausible.

We can extend the notion of what it means to be plausible by referring to the work of Kaplan and Craver (2011). The authors posit that models with explanatory power provide a "model-to-mechanism-mapping" (3M):

In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces,

maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism. (Kaplan & Craver, 2011, p. 611)

With this in mind, being plausible means to provide a good mapping of model components to the mechanism of fixation selection in the human brain. Deciding between equally predictive models therefore amounts to comparing model-to-mechanism-mappings. Conversely, a bad mapping also provides a rationale to discard models as good explanations even before we have evaluated their predictive power.

Given the importance of the model-to-mechanism-mapping, we can now ask whether the saliency map model of attention can be mapped to the mechanism of fixation selection. Attempting to provide an answer to this question confronts us with the problem that we have only incomplete knowledge of the mechanism that we want to explain. The studies presented in chapters 3-5, contribute to our understanding of the fixation selection mechanism, and therefore to our ability to generate a mapping.

Chapter 3 investigated temporal and spatial properties of eye movement trajectories. In particular, we found that previously fixated locations were often revisited. This speaks against a causal role of inhibition of return for the fixation selection process. In turn, this makes it highly unlikely that the inhibition of return component of the saliency map model can be mapped to a part of the oculomotor control mechanism in the brain.

This leads to a more general problem with the saliency map model of attention. Saliency models are usually evaluated by directly comparing the saliency map to where observers fixated. The rationale is that more salient locations should be fixated more often. The actual saccade generation process, i.e. inhibition of return, is thus not part of the evaluation process. This leads to the somewhat perplexing situation that saliency map models are performing quite

well at predicting saccadic endpoints without actually providing a mechanism to generate saccades. Of course, removing the saccade generation process from the model makes it impossible to provide a complete model-to-mechanisms-mapping. One way to improve the model, is therefore to incorporate a more realistic saccade generation mechanism.

These considerations bring up the question whether a time-invariant saliency map can plausibly be a part of predicting fixation locations. In previous work (Wilming, Betz, Harst, Waterkamp, & König, 2011)[2] we evaluated a model with a dynamical and stochastic fixation selection process (see Figure 7.1 for more information). In short, our model combined a time-invariant saliency map with time-variant spatial and saccade direction biases. The model generates predictions by estimating the probability of saccading to a location conditional on parts of the past saccade trajectory. We generated sequences of eye movements by repeatedly sampling from this distribution, which was updated after each saccade. The model therefore generated time-variant predictions of fixation probability based on a static saliency map. We found that the model generates more realistic eye movement trajectories (e.g. saccade lengths) compared to inhibition of return, while retaining the predictive power of the static saliency map. Of course, whether or not such an approach can also plausibly be part of a model-to-mechanism-mapping needs to be investigated by further studies.

The saccadic momentum effect demonstrated in chapter 3 shows that the decision of where to fixate next is dependent on the last fixation location. It is thus a prime example for the dynamic nature of the fixation selection process. It is tempting to test whether or not changes in fixation durations described by saccadic momentum can be explained by dynamic models of viewing behavior. In unpublished work (Harst, 2013), we have used an ideal-observer model (Najemnik & Geisler, 2005; Najemnik & Geisler, 2008) of eye

[2]Download here: http://ikw.uni-osnabrueck.de/~nwilming/occam_poster.png or http://ikw.uni-osnabrueck.de/~nwilming/occam_poster-web.pdf

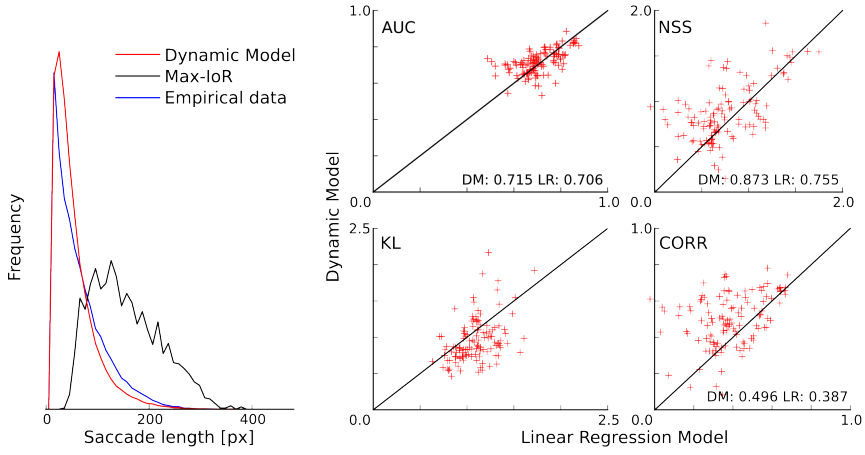


Figure 7.1: Results of modeling saccade generation with a dynamical stochastic model (adapted from Wilming, Betz, Harst, Waterkamp, & König, 2011). The dynamical model samples trajectories by combining saliency, a time-varying spatial bias and a time-and-space varying saccade bias. Fixation locations are then predicted by pooling many generated trajectories. Importantly, the dynamical model allows saccades to return to previously fixated locations. The "Max-IoR" model generates trajectories by saccading from peak to peak on the saliency map by inhibiting previously fixated locations and saccading to the maximum of the saliency map. The left histograms compare saccade lengths in empirical data with those generated with inhibition of return or the dynamical model. The dynamical model clearly produces a more realistic saccade length distribution. The four scatter plots compare how well trajectories from the dynamical model and a baseline linear regression saliency map model (same model as in Wilming, Jutras, Buffalo, & König, n.d.) predict fixation locations. Each plots shows one evaluation score (AUC, NSS, KL & correlation, c.f. Wilming, Betz, Kietzmann, & König, 2011) for the dynamical model and the baseline saliency map model. Each data point shows how well the viewing behavior of a single observer is predicted across 64 urban scenes. All four measures indicate that the predictive power of both models is comparable, suggesting that a more realistic dynamical saccade generation process is not detrimental to model performance.

movements during visual search, to investigate whether saccadic momentum could be explained by evidence accumulation of the ideal-observer. The ideal observer searches for a faint Gabor patch in a pink-noise image in an optimal fashion. With each saccade it samples evidence about the presence of the target from locations around the current fixation location. Due to the falloff of visual acuity the amount of evidence sampled about a location is dependent on the distance to the fixation location. The sampled evidence is used to update the observer's estimate of the probability (e.g. it accumulates evidence) that the target is at a given location. We found that, how much evidence is accumulated about a location depends on the angle between the past saccade and the target. Intuitively, the ideal observer samples more evidence about locations that are backwards, relative to the last saccade, compared to forward locations. Backward locations are, by definition, between the last two fixations and allow the ideal-observer to sample evidence about the location with high acuity twice. The forward target is close to the current fixation but further away from the previous location. The ideal-observer therefore has less opportunity to sample evidence about the forward target. We found that the interaction between evidence accumulation and saccade angles is reminiscent of saccadic momentum, where the fixation duration is dependent on the angle between the last and the next saccade. This is clearly suggestive of a mechanism where the duration of fixation is proportional on the evidence sampled about a location so far. But clearly, such observations can only be a starting point for further investigations.

Neither the model model presented above, nor the ideal-observer model by Najemnik and Geisler (2005) propose biological mechanisms for saccade generation. They therefore improve the model-to-mechanism-mapping only slightly. But since they emphasize the dynamical nature of the fixation selection process, it seems possible to extend these approaches to address phenomena that cannot be modeled with a static saliency map model.

Having discussed that the saliency map model does not ade-

quately address some of the phenomena in viewing behavior (the dynamic nature of fixation selection), the question remains how well we can map the rest of the saliency map model to the mechanism of fixation selection. This requires that we map the neural basis of fixation selection to the saliency map model.

In particular the introduction showed that there are multiple candidates for a saliency map in the brain. I argued that it is difficult to distinguish saliency computing areas from saliency modulated areas in a network of connected areas (see Figure 1.5). The study in chapter 4 aimed to address this problem for early visual areas. We found that early visual areas likely do not compute saliency. Instead, our results suggest that V1's role could be that of a feature map in the saliency map model of attention.

This suggestion is at odds with studies that demonstrate saliency effects in V1 (Zhaoping & May, 2007; Zhang et al., 2012). This discrepancy can be resolved when we consider how a feature map would respond to the simple stimuli (e.g. arrays of oriented bars) used in these studies. For demonstration purposes, I picked two stimuli that Li (2002, Figure 2a,b) used to exemplify the computation of saliency in V1. The two stimuli demonstrate a visual search asymmetry: A cross among bars is easier to find (i.e. the unique cross is salient) than a bar among crosses (i.e. the bar is not salient, even though it is unique). Li Zhaoping's V1 model produces higher activation for a cross among bars than for a bar among crosses. The V1 model therefore "explains" the difference in saliency between the two stimuli. Yet, computing a Gabor feature map on the same stimuli also produces the search asymmetry (Figure 7.2). Importantly, Gabor kernels are often used as simple features in saliency map models (for example in Itti & Koch, 2000). Thus, in this specific example, we cannot differentiate between a saliency map and a feature (or conspicuity) map.

The problem is more general though. Saliency maps integrate over many conspicuity maps. Conspicuity maps, by definition, contain high scores for elements that are distinct from their surrounding.

They therefore correlate with saliency to some extend. Distinguishing between saliency and conspicuity maps requires stimuli where saliency and feature computation are dissociated. In chapter 4 we have implemented this with high and low contrast manipulations, such that luminance contrast did not linearly correlate with saliency. An alternative approach would be to use stimuli where locations are salient because conspicuity of different features adds up, but each individual conspicuity map is more or less homogeneous.

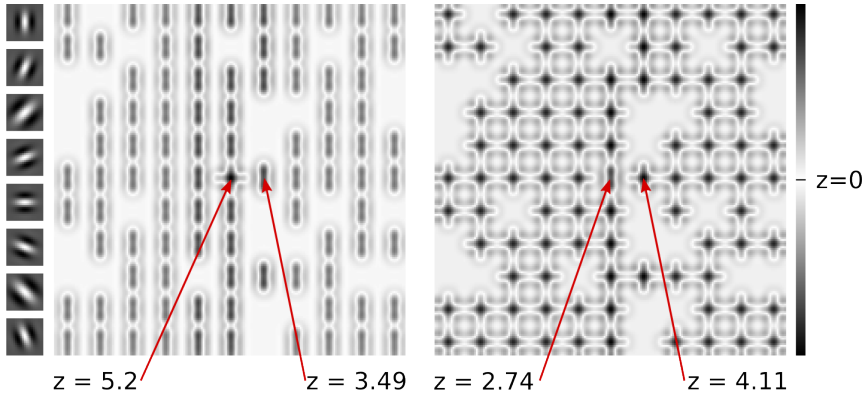


Figure 7.2: Gabor features and saliency on simple stimuli. The left column shows a set of oriented Gabor filters that were used to simulate simple cells by convolving the filters with the stimuli. The remaining panels show the summed filter response (i.e. a conspicuity map) to two stimuli used in Li (2002). The original stimuli (Figure 2a,b Li, 2002) demonstrate a visual search assymetry: A cross among bars is easier to find than a bar among crosses. The z values below the filter output give the z-score of the cross and the bar, showing that a Gabor conspicuity map assigns higher activity to a cross among bars and lower activity to a bar among crosses. The Gabor conspicuity map therefore "explains" the search asymmetry.

These considerations stress the importance for establishing a model-to-mechanism-mapping that includes all candidate areas for saliency maps. The study presented in chapter 4 asserts that V1 can likely not be mapped to a saliency map. In turn, this makes it more likely that one of the other areas reviewed in the introduction is a gen-

uine saliency map. At least, our results suggest that they do not simply receive saliency modulated feed-forward activity from V1. Providing a model-to-mechanism-mapping is an endeavor that will require further work to clarify the role of many areas that have been implicated in the processing of saliency.

The manuscript in chapter 5 pertains to this aspect by suggesting that the guidance of eye movements by saliency is comparable in monkeys and humans. This, of course, does not show that oculomotor areas in the macaque brain have 1:1 homologues in the human brain. In so far results in chapter 5 provide no direct constraints for a model-to-mechanism-mapping. However, the results provide a very detailed examination of the effect of saliency during free viewing of pictures in monkeys. This has important consequences for electrophysiological studies of saliency that use tasks with free eye movements.

For example, in an excellent study, Fernandes, Stevenson, Phillips, Segraves, and Körding (2013) investigate neural processing during search for a picture of a fly embedded in a natural scene. The authors employ a generative model that combines saliency with saccade direction to predict firing behavior of neurons in FEF. They find that saliency explains only a tiny fraction of FEF spikes. This would suggest that FEF is not a saliency map. However, several aspects make the interpretation of this conclusion difficult. First, the background scene that is used to compute saliency is completely irrelevant for the task. Ignoring saliency might therefore be beneficial for solving the task. Second, it is known that FEF integrates top-down information and thus encodes more than saliency. These considerations might explain why the authors find that saliency is only weakly predictive of fixation locations during their task (AUC 0.58; our study: AUC 0.69 for urban scenes). In summary, it seems that free viewing is a much fairer task for investigating saliency. A contribution of chapter 5 is therefore the assertion that studying the effect of saliency in monkeys is feasible with a free viewing task.

As presented above, the results in this thesis have important impli-

cations for establishing a model-to-mechanism-mapping. An issue that this thesis has not touched upon is the question why our viewing behavior is the way it is. Assuming that the brain uses saliency to direct attention, how would we explain why an observer fixated an object and not another? We would probably try to measure activity in the saliency map, trace this activity through all conspicuity maps and conclude that one object was more salient because of its constituent features. But this explanations has nothing to say about why our oculomotor control system selects fixation location from a saliency map. As such, the saliency map model of attention does not have a normative component.

In the last chapter we have commented on an article by Clark (2013). Clark makes a compelling case for a normative model of cognition. In this model, different levels in the brain predict the activity of lower levels and thereby match their "expectations" to the sensory input. The goal of the brain is to infer the causes of incoming sensory signals. How well expectations match sensory signals now becomes a measure of how well the brain can infer the causal structure of its environment. In this view, applying the same prediction machinery hierarchically supposedly leads to more and more abstract models for causes in the environment (e.g. from simple cells in V1 to object cells in IT). Such a view suggests that we can understand large parts of cortical processing in the normative light of predictive coding. One example of predictive coding is the slowness principle (Creutzig & Sprekeler, 2008). Tuning of neurons towards slowly varying properties of the environment leads to the development of simple and complex cells (Berkes & Wiskott, 2005; Einhäuser, Kayser, König, & Körding, 2002b), and, in combination with sparse activation of units, to place, head-direction and spatial-view cells (Wyss, Konig, Verschure, & König, 2006; Franzius, Sprekeler, & Wiskott, 2007).

What are the consequences of using normative models for understanding overt and covert attention mechanisms? Two different aspects appear especially important. First, the fact that normative principles can, to some extend, explain neural processing in some

brain areas suggests that this improves our understanding of these areas. An improved understanding of neural computations should lead to direct improvements of attention models since these potentially depend on the same computations. Second, Feldman and Friston (2010) discuss a predictive coding model where perception is inference about causes of sensory signals and attention is inference about the uncertainty of those causes. In this scheme covert attention is high when the certainty about the cause of sensory signals is high. Friston, Adams, Perrinet, and Breakspear (2012) take the same principles one step further to model eye movements. In this proposal, hypotheses about sensory causes are modeled in tandem with eye movements that would test these hypotheses. Here, perception and action form a continuous cycle. Perception infers potential causes of the sensory input and actions generate sensory input that confirms or rejects hypotheses about potential causes. A proposal that strongly resonates with embodied views of cognition (just to name one: O'Regan & Nöe, 2001).

Whether or not models derived from such principles can provide satisfying explanations of oculomotor control is, of course, a question of ongoing debate. But their emergence raises questions about the role of saliency which deserve further investigations. Independent of the outcome - the possibilities offered by such models are thrilling.

The results from this thesis provide different perspectives on how future explanations of viewing behavior might look like. First, we suggest that unraveling the mechanism of oculomotor control requires computational models that are more than good predictors of viewing behavior. Models need to be rigorously mapped to the mechanism of fixation selection. Second, viewing behavior is a spatio-temporal decision process and modelers must treat it as such. Future models need to embrace the spatio-temporal structure of eye-movements to provide valuable insights into the fixation selection process. Third, the model-to-mechanism mapping will likely not associate early visual areas with a saliency map. Fourth, there

is now evidence for a strong similarity between humans and monkeys in saliency-driven viewing behavior. The saliency-processing areas that have been identified in monkeys therefore likely figure in model-to-mechanism mappings. Taking these recommendations into account might lead to new models of oculomotor control that provide explanations of the fixation selection mechanism. And this will ultimately help to answer the much deeper question of how our mind establishes contact with the world.

Acknowledgements

This thesis would not have been possible without the help of many others. First and foremost I'd like to thank my PhD supervisors who have supported me in the best possible way. Peter, thank you for your invaluable support, countless opportunities and many lessons learned. But most of all, thank you for having such an optimistic attitude towards life! Beth, coming to Atlanta and Seattle meant a lot to me. Thank you for giving me this huge opportunity, thank you for your support, for your advice and your encouragement!

I'd furthermore like to thank my co-authors for their invaluable help and support. I'd like to especially thank John-Dylan Haynes and Carsten Bogler for supporting the fMRI measurements in Berlin and Leipzig. Torsten, Tim, Nico, Simon and Megan: Cheers to you! Let's do this again!

I've had the pleasure of working in two amazing labs. Dear NBP, thank y'all for providing such a welcoming and open environment. Working with you was and is a pleasure! Dear Buffalo lab, thank y'all for lively discussions and welcoming a complete stranger into the lab. Megan thank you for being such a strong support for a complete greenhorn.

Many friends shared parts of the way with me. Torsten, Tim, Robert, Cornell and Jose: Thank you for everything! This thesis would not exist without you.

Torsten, Jose, Johannes and Robert have proof-read parts of this thesis and found countless mistakes. Thank you!

The IKW happens to have some of the best admins in the world. You guys do an awesome job!
Dear family: You are the best!
Paula and Sarah, this thesis is for you.

Bibliography

Açik, A., Onat, S., Schumann, F., Einhäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research*, *49*(12), 1541–53.

Açik, A., Sarwary, A., Schultze-Kraft, R., Onat, S., & König, P. (2010). Developmental Changes in Natural Viewing Behavior: Bottom-Up and Top-Down Differences between Children, Young Adults and Older Adults. *Frontiers in psychology*, *1*, 207.

Amunts, K., Malikovic, a., Mohlberg, H., Schormann, T., & Zilles, K. (2000). Brodmann's areas 17 and 18 brought into stereotaxic space-where and how variable? *NeuroImage*, *11*(1), 66–84.

Anderson, A. J., Yadav, H., & Carpenter, R. H. S. (2008). Directional prediction by the saccadic system. *Current Biology*, *18*(8), 614–618.

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

Arcizet, F., Mirpour, K., & Bisley, J. W. (2011a). A pure salience response in posterior parietal cortex. *Cerebral cortex*, *21*(11), 2498–506.

Arcizet, F., Mirpour, K., & Bisley, J. W. (2011b). A pure salience response in posterior parietal cortex. *Cerebral cortex (New York, N.Y. : 1991)*, *21*(11), 2498–506.

Baddeley, R. J. & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, *46*(18), 2824–33.

Bahill, A. T., Clark, M. R., & Stark, L. (1975). The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, *204*, 191–204.

Bahill, A. & LaRitz, T. (1984). Why can't batters keep their eyes on the ball? *American Scientist*, *72*(3), 249–253.

Ballard, D. & Hayhoe, M. (1992). Hand-eye coordination during sequential tasks. *Phil. Trans. R. Soc. Lond. B*, *337*(1281), 331–339.

Barlow, H. B. H. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, 217–234.

Basso, M. a. & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *Journal of Neuroscience*, *18*(18), 7519–34.

Bastos, A., Vezoli, J., & Bosman, C. (2014). Visual areas exert feedforward and feedback influences through distinct frequency channels. *bioRxiv*.

Bays, P. & Husain, M. (2012). Active inhibition and memory promote exploration and search of natural scenes. *Journal of Vision*, *12*, 1–18.

Bell, A. H. & Munoz, D. P. (2008). Activity in the superior colliculus reflects dynamic interactions between voluntary and involuntary influences on orienting behaviour. *The European Journal of Neuroscience*, *28*(8), 1654–60.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, *9*(5), 1–15.

Berkes, P. & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 579–602.

Betsch, B. Y., Einhäuser, W., Körding, K. P., & König, P. (2004). The world from a cat's perspective: statistics of natural videos. *Biological Cybernetics*, *90*(1), 41–50.

Betz, T., Kietzmann, T. C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, *10*, 1–14.

Betz, T., Wilming, N., Bogler, C., Haynes, J.-D., & König, P. (2013). Dissociation between saliency signals and activity in early visual cortex. *Journal of Vision*, *13*(14), 1–12.

Bichot, N. P., Thompson, K. G., Chenchal Rao, S., & Schall, J. D. (2001). Reliability of macaque frontal eye field neurons signaling saccade targets during visual search. *Journal of Neuroscience*, *21*(2), 713–25.

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*(5721), 529–34.

Bichot, N. P. & Schall, J. D. (2002). Priming in macaque frontal cortex during popout visual search: feature-based facilitation and location-based inhibition of return. *Journal of Neuroscience*, *22*(11), 4675–85.

Bichot, N., Schall, J. D., & Thompson, K. (1996). Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature*, *381*(20), 697.

Bisley, J. W. (2011). The neural basis of visual attention. *The Journal of physiology*, *589*(Pt 1), 49–57.

Bisley, J. W. & Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science*, *299*(5603), 81–6.

Bisley, J. W. & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review Of Neuroscience*, *33*, 1–21.

Blatt, G. (1990). Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. *Journal of Comparative Neurology*, *299*, 421–445.

Bogler, C., Bode, S., & Haynes, J.-D. (2011). Decoding Successive Computational Stages of Saliency Processing. *Current Biology*, 1–5.

Borji, A. & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, *91*, 62–77.

Bridgeman, B., Hendry, D., & Stark, L. (1975). Failure to detect displacement of the visual world during saccadic eye movements. *Vision Research*, *15*(6), 719–22.

Brown, J. M. & Guenther, B. a. (2012). Magnocellular and parvocellular pathway influences on location-based inhibition-of-return. *Perception*, *41*(3), 319–338.

Bruce, N. & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems*.

Bruce, N. & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 1–24.

Buffalo, E. a., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 361–5.

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

Buschman, T. J. & Miller, E. K. (2009). Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. *Neuron*, *63*(3), 386–96.

Butko, N. J. & Movellan, J. R. (2008). I-POMDP: An infomax model of eye movement. *2008 7th IEEE International Conference on Development and Learning*, (1), 139–144.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., … Rust, N. C. (2005). Do we know what the

early visual system does? *Journal of Neuroscience*, *25*(46), 10577–97.

Carello, C. D. & Krauzlis, R. J. (2004). Manipulating intent: evidence for a causal role of the superior colliculus in target selection. *Neuron*, *43*(4), 575–83.

Castelhano, M., Mack, M., & Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*, 1–15.

Cavanaugh, J., Alvarez, B. D., & Wurtz, R. H. (2006). Enhanced performance with brain stimulation: attentional shift or visual cue? *Journal of Neuroscience*, *26*(44), 11347–58.

Cavanaugh, J. & Wurtz, R. H. (2004). Subcortical modulation of attention counters change blindness. *Journal of Neuroscience*, *24*(50), 11236–43.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task : Experimental data and computer model. *Journal of Vision*, *9*, 1–15.

Cerf, M., Harel, J., Einauser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, *20*, 241–248.

Chao, A. & Shen, T. J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*.

Chauvin, a. & Herault, J. (2002). Natural scene perception: visual attractors and images processing. *Progress in Neural Processing*. Progress in Neural Processing, 236–248.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., … Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron*, *21*(4), 761–73.

Craver, C. F. (2007). Explaining the Brain Mechanisms and the Mosaic Unity of Neuroscience. In *Explaining the brain: mechanisms and the mosaic unity of neuroscience* (pp. 229–271).

Creutzig, F. & Sprekeler, H. (2008). Predictive coding and the slowness principle: an information-theoretic approach. *Neural Computation*, *20*(4), 1026–41.

Curcio, C., Sloan, K., Packer, O., Hendrickson, A., & Kalina, R. (1987). Distribution of cones in human and monkey retina: individual variability and radial asymmetry. *Science*, (10), 579.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *194*, 179–194.

David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. *Neuron*, *59*, 509–521.

Deubel, H. & Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–37.

Diamond, M. E., von Heimendahl, M., Knutsen, P. M., Kleinfeld, D., & Ahissar, E. (2008). 'Where' and 'what' in the whisker sensorimotor system. *Nature reviews. Neuroscience*, *9*(8), 601–12.

Dorris, M. C. & Munoz, D. P. (1998). Saccadic probability influences motor preparation signals and time to saccadic initiation. *Journal of Neuroscience*, *18*(17), 7015–26.

Dorris, M. C., Klein, R. M., Everling, S., & Munoz, D. P. (2002). Contribution of the primate superior colliculus to inhibition of return. *Journal of cognitive neuroscience*, *14*(8), 1256–1263.

Dorris, M. C., Olivier, E., & Munoz, D. P. (2007). Competitive integration of visual and preparatory signals in the superior colliculus during saccadic programming. *Journal of Neuroscience*, *27*(19), 5053–62.

Ehinger, K. K. a., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: a combined source model of eye guidance. *Visual cognition*, *17*(1935), 1–30.

Eickhoff, S., Stephan, K., Mohlberg, H., Grefkes, C., Fink, G., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, (4), 1325–1335.

Einhäuser, W. & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, *17*(5), 1089–1097.

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002a). Learning the invariance properties of complex cells from their responses to natural stimuli. *The European Journal of Neuroscience*, *15*(3), 475–86.

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002b). Learning the invariance properties of complex cells from their responses to natural stimuli. *The European Journal of Neuroscience*, *15*(3), 475–86.

Einhäuser, W., Kruse, W., Hoffmann, K.-P., & König, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, *46*(8-9), 1194–209.

Einhäuser, W., Moeller, G. U., Schumann, F., Conradt, J., Vockeroth, J., Bartl, K., … König, P. (2009). Eye-head coordination during free exploration in human and cat. *Annals of the New York Academy of Sciences*, *1164*, 353–366.

Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*(11), 1148–58.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*.

Elazary, L. & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 31–15.

Engbert, R. & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045.

Erdem, E. & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, *13*(4), 1–20.

Farrell, S., Ludwig, C. J. H., Ellis, L. a., & Gilchrist, I. D. (2010). Influence of environmental statistics on inhibition of saccadic return. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 929–934.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fecteau, J. H. & Munoz, D. P. (2005). Correlates of capture of attention and inhibition of return across stages of visual processing. *Journal of cognitive neuroscience*, *17*(11), 1714–27.

Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, *4*, 215.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1–47.

Fernandes, H. L., Stevenson, I. H., Phillips, a. N., Segraves, M. a., & Körding, K. P. (2013). Saliency and Saccade Encoding in the Frontal Eye Field During Natural Scene Search. *Cerebral Cortex*, 1–14.

Fischer, J. & Whitney, D. (2012). Attention gates visual coding in the human pulvinar. *Nature communications*, *3*, 1051.

Flanagan, J. R. & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–71.

Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, *3*(8), e166.

Frey, H.-P., Wirz, K., Willenbockel, V., Betz, T., Schreiber, C., Troscianko, T., & König, P. (2011). Beyond correlation: do color features influence attention in rainforest? *Frontiers in human neuroscience*, *5*(April), 36.

Frey, H., Honey, C., & König, P. (2008). What ' s color got to do with it ? The influence of color on visual attention in different categories. *Journal of Vision*, *8*(14), 1–17.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, *11*, 127–138.

Friston, K., Adams, R. a., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, *3*(May), 151.

Fujii, N., Mushiake, H., & Tanji, J. (1998). Intracortical microstimulation of bilateral frontal eye field. *Journal of neurophysiology*, *79*(4), 2240–4.

Gandhi, N. J. & Katnani, H. a. (2011). Motor functions of the superior colliculus. *Annual Review Of Neuroscience*, *34*, 205–31.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, *8*(7), 1–18.

Gao, D. & Vasconcelos, N. (2004). Discriminant Saliency for Visual Recognition from Cluttered Scenes. *NIPS*.

Gardner, J. L., Sun, P., Waggoner, R. A., Ueno, K., Tanaka, K., & Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. *Neuron*, *47*(4), 607–20.

Gattass, R. & Desimone, R. (2014). Effect of Microstimulation of the Superior Colliculus on Visual Space Attention. *Journal of cognitive neuroscience*, 1–12.

Gattass, R., Galkin, T. W., Desimone, R., & Ungerleider, L. G. (2014). Subcortical connections of area V4 in the macaque. *The Journal of Comparative Neurology*, *522*(8), 1941–65.

Gaymard, B., Lynch, J., Ploner, C., Condy, C., & Rivaud-Pechoux, S. (2003). The parieto-collicular pathway: anatomical location and contribution to saccade generation. *European Journal of Neuroscience*, *17*(7), 1518–1526.

Gelman, A. & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331.

Geng, J. J. & Mangun, G. R. (2009). Anterior intraparietal sulcus is sensitive to bottom-up attention driven by stimulus salience. *Journal of cognitive neuroscience*, *21*(8), 1584–1601.

Ghazanfar, A. a., Nielsen, K., & Logothetis, N. K. (2006). Eye movements of monkey observers viewing vocalizing conspecifics. *Cognition*, *101*(3), 515–29.

Gilbert, C. D. & Li, W. (2013). Top-down influences on visual processing. *Nature reviews. Neuroscience*, *14*(5), 350–63.

Gilland, J. (2008). *Driving, eye-tracking and visual entropy: Exploration of age and task effects* (Doctoral dissertation, The University of South Dakota).

Glimcher, P. W. & Sparks, D. L. (1992). Movement selection in advance of action in the superior colliculus. *Nature*, *355*(6360), 542–5.

Goddard, E., Mannion, D. J., McDonald, J. S., Solomon, S. G., & Clifford, C. W. G. (2011). Color responsiveness argues against a dorsal component of human V4. *Journal of Vision*, *11*(4), 1–21.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*(6666), 481–484.

Grieve, K. L., Acuña, C., & Cudeiro, J. (2000). The primate pulvinar nuclei: vision and action. *Trends in neurosciences*, *23*(1), 35–9.

Guo, K., Meints, K., Hall, C., Hall, S., & Mills, D. (2009). Left gaze bias in humans, rhesus monkeys and domestic dogs. *Animal cognition*, *12*(3), 409–18.

Guo, K., Robertson, R. G., Mahmoodi, S., Tadmor, Y., & Young, M. P. (2003). How do monkeys view faces?–A study of eye movements. *Experimental brain research. 150*(3), 363–74.

Hafed, Z. M. & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research*, *42*(22), 2533–45.

Hafed, Z. M. & Krauzlis, R. J. (2006). Ongoing eye movements constrain visual perception. *Nature Neuroscience*, *9*(11), 1449–57.

Hall, N. J. & Colby, C. L. (2011). Remapping for visual stability. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *366*(1564), 528–39.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*.

Harst, S. (2013). *Is Information from previous fixations incorporated into a prior of bayesian-optimal search behavior?* (Master thesis, University of Osnabrück).

Harting, J. K., Huerta, M., Frankfurter, A., Strominger, N., & Royce, G. (1980). Ascending Pathways From the Monkey Superior Colliculus: An Autoradiographic Analysis. *The Journal of Comparative Neurology*, *882*(192), 853–882.

Hausser, J. & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning*, *10*, 1469–1484.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition*, *7*(1997), 43–64.

Hayhoe, M. M., McKinney, T., Chajka, K., & Pelz, J. B. (2012). Predictive eye movements in natural vision. *Experimental brain research*, *217*(1), 125–36.

Haynes, J.-D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature reviews. Neuroscience*, *7*(7), 523–34.

Heinzle, J., Kahnt, T., & Haynes, J.-D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. *NeuroImage*, *56*(3), 1426–1436.

Henderson, J., Brockmole, J., Castelhano, M., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements: a window on mind and brain* (pp. 537–562).

Hoffman, J. (1998). Visual attention and eye movements. In *Attention* (pp. 121–153).

Holste, D., Grosse, I., & Herzel, H. (1998). Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General*, *31*, 2551.

Hooge, I. T. & Frens, M. a. (2000). Inhibition of saccade return (ISR): spatio-temporal properties of saccade programming. *Vision Research*, *40*(24), 3415–3426.

Hooge, I. T. C., Over, E. a. B., van Wezel, R. J. a., & Frens, M. a. (2005). Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research*, *45*(14), 1901–8.

Horwitz, G. D. & Newsome, W. T. (2001). Target selection for saccadic eye movements: prelude activity in the superior colliculus during a direction-discrimination task. *Journal of neurophysiology*, *86*(5), 2543–58.

Horwitz, G. D., Batista, A. P., & Newsome, W. T. (2004). Representation of an abstract perceptual decision in macaque superior colliculus. *Journal of neurophysiology*, *91*(5), 2281–96.

Hubel, D. & Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas. *Journal of Neurophysiology*, *28*(2), 229.

Huerta, M., Krubitzer, L., & Kaas, J. (1986). Frontal eye field as defined by intracortical microstimulation in squirrel monkeys, owl monkeys, and macaque monkeys: I. Subcortical connections. *Journal of Comparative … 439*, 415–439.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–6.

Hwang, A. A., Higgins, E. E. E., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5), 1–18.

Ipata, A. E., Gee, A. L., Bisley, J. W., & Goldberg, M. E. (2009). Neurons in the lateral intraparietal area create a priority map by the combination of disparate signals. *Experimental brain research*, *192*(3), 479–88.

Ipata, A. E., Gee, A. L., & Goldberg, M. E. (2012). Feature attention evokes task-specific pattern selectivity in V4 neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), 16778–85.

Ipata, A. E., Gee, A. L., Goldberg, M. E., & Bisley, J. W. (2006). Activity in the lateral intraparietal area predicts the goal and latency of saccades in a free-viewing visual search task. *Journal of Neuroscience*, *26*(14), 3656–61.

Ipata, A. E., Gee, A. L., Gottlieb, J., Bisley, J. W., & Goldberg, M. E. (2006). LIP responses to a popout stimulus are reduced if it is overtly ignored. *Nature Neuroscience*, *9*(8), 1071–6.

Itti, L. & Baldi, P. (2005a). A principled approach to detecting surprising events in video. In *Ieee computer society conference on computer vision and pattern recognition* (pp. 631–637).

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

Itti, L. & Koch, C. (2001a). Computational modelling of visual attention. *Nat Rev Neurosci*, *2*(3), 194–203.

Itti, L. & Koch, C. (2001b). Computational modelling of visual attention. *Nature reviews. Neuroscience*, *2*(3), 194–203.

Itti, L. & Baldi, P. (2005b). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–306.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

James, W. (1890). *The principles of psychology, Vol I.*

Jansen, L., Onat, S., & König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, *9*(1), 1–19.

Jerde, T. a., Merriam, E. P., Riggall, A. C., Hedges, J. H., & Curtis, C. E. (2012). Prioritized maps of space in human frontoparietal cortex. *Journal of Neuroscience*, *32*(48), 17382–90.

Johnson, L., Sullivan, B., Hayhoe, M., & Ballard, D. (2014). Predicting human visuomotor behaviour in a driving task. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1636), 20130044.

Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations*.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *Computer Vision*.

Jutras, M. J. & Buffalo, E. a. (2010). Recognition memory signals in the macaque hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 401–6.

Jutras, M. J., Fries, P., & Buffalo, E. a. (2009). Gamma-band synchronization in the macaque hippocampus and memory formation. *Journal of Neuroscience*, *29*(40), 12521–31.

Kagan, I., Iyer, A., Lindner, A., & Andersen, R. a. (2010). Space representation for eye movements is more contralateral in monkeys than in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7933–8.

Kanan, C., Tong, M., Zhang, L., & Cottrell, G. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition*, *17*(6), 979–1003.

Kano, F. & Tomonaga, M. (2009). How chimpanzees look at pictures: a comparative eye-tracking study. *Proceedings. Biological sciences / The Royal Society*, *276*(1664), 1949–55.

Kaplan, D. & Craver, C. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*. *Philosophy of Science*, *78*(October), 601–627.

Kärcher, S. M., Fenzlaff, S., Hartmann, D., Nagel, S. K., & König, P. (2012). Sensory Augmentation for the Blind.

Kaspar, K. & König, P. (2011a). Overt attention and context factors: the impact of repeated presentations, image type, and individual motivation. *PLoS ONE*, *6*(7), e21719.

Kaspar, K. & König, P. (2011b). Viewing behavior and the impact of low-level image properties across repeated presentations of complex scenes. *Journal of Vision*, *11*, 1–29.

Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: interaction of image structure and image content. *Vision Research*, *46*(16), 2535–45.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5), 1–15.

Kienzle, W., Wichmann, F., & Scholkopf, B. (2007). A Nonparametric Approach to Bottom-Up Visual Saliency. *Advances in Neural Information Processing Systems*, 689–696.

Kietzmann, T., Geuter, S., & König, P. (2011). Overt Visual Attention as a Causal Factor of Perceptual Awareness. *PLoS ONE*, *6*(7).

Klein, R. & MacInnes, W. (1999). Inhibition of return is a foraging fascilitator in visual search. *Psychological science*, *10*(4), 346–352.

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry.

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology*, *6*(5), e1000791.

König, P. & Krüger, N. (2006). Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics*, *94*(4), 325–334.

König, P., Wilming, N., Kaspar, K., Nagel, S. K., & Onat, S. (2013). Predictions in the light of your own action repertoire as a general computational principle. *Behavioral and Brain Sciences*, *36*(03), 39–40.

Kootstra, G., de Boer, B., & Schomaker, L. R. B. (2011). Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry. *Cognitive Computation*, *3*(1), 223–240.

Kootstra, G., Wilming, N., Schmidt, N., Djurfeldt, M., Kragic, D., & Peter, K. (2012). Learning and Adaptation of Sensorimotor Contingencies : Prism-Adaptation, a case study. In *From animals to animats 12 - 12th international conference on simulation of adaptive behavior proceedings.* (Vol. 7426/2012, pp. 341–350). Springer.

Körding, K. P., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *Journal of neurophysiology*, *91*(1), 206–12.

Kovesi, P. (1999). Image features from phase congruency. *Videre: Journal of Computer Vision Research*, *1*(3), 1–26.

Kovesi, P. (2003). Phase congruency detects corners and edges. In *The australian pattern recognition society conference* (pp. 309–318). Sydney, USA: IEEE.

Krauzlis, R. J., Lovejoy, L. P., & Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annual Review Of Neuroscience*, *36*, 165–82.

Krauzlis, R. & Dill, N. (2002). Neural correlates of target choice for pursuit and saccades in the primate superior colliculus. *Neuron*, *35*(2), 355–63.

Krichevsky, R. & Trofimov, V. (2002). The performance of universal encoding. *IEEE Transactions on Information Theory*, *27*(2), 199–207.

Kriegeskorte, N., Mur, M., Ruff, D., & Kiani, R. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868.

Kuang, X., Poletti, M., Victor, J. D., & Rucci, M. (2012). Temporal encoding of spatial information during active visual fixation. *Current Biology*, *22*(6), 510–4.

Künzle, H., Akert, K., & Wurtz, R. H. (1976). Projection of area 8 (frontal eye field) to superior colliculus inthe monkey. An autoradiographic study. *Brain Research*, *117*(Pt 10), 487–492.

Kustov, A. A. & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, *384*, 74–77.

LaBerge, D. & Buchsbaum, M. (1990). Positron emission tomographic measurements of pulvinar activity during an attention task. *Journal of Neuroscience*, *10*(February), 613–619.

Land, M. F. & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Phil. Trans. R. Soc. Lond. B*, *352*(1358), 1231–9.

Land, M. F. & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25-26), 3559–65.

Land, M. F. & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, *3*(12), 1340–5.

Land, M. F. & Tatler, B. W. (2001). Steering with the head. the visual strategy of a racing driver. *Current Biology*, *11*(15), 1215–20.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311–1328.

Laubrock, J., Engbert, R., & Kliegl, R. (2005). Microsaccade dynamics during covert attention. *Vision Research*, *45*(6), 721–30.

Li, Z. (1998). A neural model of contour integration in the primary visual cortex. *Neural Computation*, *10*(4), 903–40.

Li, Z. (1999a). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(18), 10530–5.

Li, Z. (1999b). Visual segmentation by contextual influences via intracortical interactions in the primary visual cortex. *Network*, *10*(2), 187–212.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9–16.

Li, Z. & Zhaoping, L. (2005). The primary visual cortex creates a bottom-up saliency map. *Neurobiology of attention*, (May).

Lovejoy, L. P. & Krauzlis, R. J. (2010). Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. *Nature Neuroscience*, *13*(2), 261–6.

Ludwig, C. J. H., Farell, S., Ellis, L. a., Gilchrist, I. D., & Farrell, S. (2009). The mechanism underlying inhibition of saccadic return. *Cognitive Psychology*, *59*(2), 180–202.

Lynch, J. C. & Tian, J.-R. (2006). Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in brain research*, *151*, 461–501.

MacKay, D. M. (1973). Visual Stability and Voluntary Eye Movements. In R. Jung (Ed.), *Handbook of sensory physiology. central processing of visual information a.* (pp. 307–331). Springer Berlin Heidelberg.

Mantini, D., Corbetta, M., Romani, G. L., Orban, G. a., & Vanduffel, W. (2012). Data-driven analysis of analogous brain networks in monkeys and humans during natural vision. *NeuroImage*, *63*(3), 1107–18.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., … Kennedy, H. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *The Journal of Comparative Neurology*, *522*(1), 225–59.

Martinez-Trujillo, J. & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*, 744–751.

Martínez-Trujillo, J. & Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, *35*(2), 365–70.

Masciocchi, C., Mihalas, S., Parkhurst, D., & Nierbur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, *9*(11), 1–22.

Mazer, J. a. & Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron*, *40*(6), 1241–50.

McAdams, C. J. & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, *19*(1), 431–41.

McFarland, R., Roebuck, H., Yan, Y., Majolo, B., Li, W., & Guo, K. (2013). Social Interactions through the Eyes of Macaques and Humans. *PLoS ONE*, *8*(2), e56437.

McPeek, R. M. & Keller, E. L. (2002). Saccade target selection in the superior colliculus during a visual search task. *Journal of neurophysiology*, *88*(4), 2019–34.

Melloni, L., van Leeuwen, S., Alink, A., & Müller, N. G. (2012). Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cerebral cortex*, *22*(12), 2943–52.

Miller, G. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods II-B*, *2*, 95–100.

Mirpour, K., Arcizet, F., Ong, W. S., & Bisley, J. W. (2009). Been there, seen that: a neural mechanism for performing efficient visual search. *Journal of neurophysiology*, *102*(6), 3481–91.

Mirpour, K. & Bisley, J. W. (2012). Dissociating activity in the lateral intraparietal area from value using a visual foraging task. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(25), 10083–8.

Monosov, I. E., Sheinberg, D. L., & Thompson, K. G. (2011). The effects of prefrontal cortex inactivation on object responses of single neurons in the inferotemporal cortex during visual search. *Journal of Neuroscience*, *31*(44), 15956–61.

Monosov, I. E. & Thompson, K. G. (2009). Frontal eye field activity enhances object identification during covert visual search. *Journal of neurophysiology*, *102*(6), 3656–72.

Moore, T. & Armstrong, K. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(January), 370–373.

Müller, J. R., Philiastides, M. G., & Newsome, W. T. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(3), 524–9.

Nagel, S. K., Carl, C., Kringe, T., Märtin, R., & König, P. (2005). Beyond sensory substitution–learning the sixth sense. *Journal of Neural Engineering*, *2*(4), R13–26.

Najemnik, J. & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3), 1–14.

Najemnik, J. & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–91.

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. *Arxiv preprint physics/0108025*.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Nobre, a. C., Gitelman, D. R., Dias, E. C., & Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: overlapping neural systems. *NeuroImage*, *11*(3), 210–6.

Nuthmann, A. & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*, 1–19.

Oleksiak, A., Postma, A., van der Ham, I. J. M., Klink, P. C., & van Wezel, R. J. a. (2011). A review of lateralization of spatial functioning in nonhuman primates. *Brain research reviews*, *67*(1-2), 56–72.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4), e93254.

Onat, S., König, P., & Jancke, D. (2011). Natural scene evoked population dynamics across cat primary visual cortex captured with voltage-sensitive dye imaging. *Cerebral cortex*, *21*(11), 2542–2554.

Orban, G. a., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–24.

O'Regan, J. K., Rensink, R. a., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature*, *398*(6722), 34.

O'Regan, J. (1992). Solving the" real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology/Revue canadienne …*

O'Regan, J. & Nöe, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, *129*(1), 79–103.

O'Shea, R. (1991). Thumb's rule tested: visual angle of thumb's width is about 2 deg. *Perception*, *20*, 415–418.

Ossandón, J. P., Onat, S., Cazzoli, D., Nyffeler, T., Müri, R., & König, P. (2012). Unmasking the contribution of low-level features to the guidance of attention. *Neuropsychologia*, *50*(14), 3478–87.

Parikh, N., Itti, L., & Weiland, J. (2010). Saliency-based image processing for retinal prostheses. *Journal of neural engineering*, *7*(1), 16006.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107±123.

Passingham, R. (2009). How good is the macaque monkey model of the human brain? *Current Opinion in Neurobiology*, *19*(1), 6–11.

Patel, G. H., Shulman, G. L., Baker, J. T., Akbudak, E., Snyder, A. Z., Snyder, L. H., & Corbetta, M. (2010). Topographic organization of macaque area LIP. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4728–33.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Petersen, S. E., Robinson, D. L., & Keys, W. (1985). Pulvinar nuclei of the behaving rhesus monkey: visual responses and their modulation. *Journal of neurophysiology*, *54*(4), 867–86.

Petersen, S. E. & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review Of Neuroscience*, *35*, 73–89.

Phillips, W., Kay, J., & Smyth, D. (1995). The discovery of structure by multi-stream networks of local processors with contextual guidance. *Neural Computation*, *6*, 225–246.

Piaget, J. (1953). The origins of intelligence in children.

Piccinini, G. & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Posner, M. I. & Cohen, Y. (1984). Components of Visual Orienting. *Attention and performance X*, *32*, 531–556.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Premereur, E., Vanduffel, W., & Janssen, P. (2011). Functional heterogeneity of macaque lateral intraparietal neurons. *Journal of Neuroscience*, *31*(34), 12307–17.

Premereur, E., Vanduffel, W., & Janssen, P. (2014). The Effect of FEF Microstimulation on the Responses of Neurons in the Lateral Intraparietal Area. *Journal of Cognitive Neuroscience*, 1–13.

Purushothaman, G., Marion, R., Li, K., & Casagrande, V. a. (2012a). Gating and control of primary visual cortex by pulvinar. *Nature Neuroscience*, *15*(6), 905–12.

Purushothaman, G., Marion, R., Li, K., & Casagrande, V. a. (2012b). Gating and control of primary visual cortex by pulvinar. *Nature Neuroscience*, *15*(6).

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rayner, K. (2009). *Eye movements and attention in reading, scene perception, and visual search.*

Recarte, M. & Nunes, L. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology Applied*, *6*(1), 31–43.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, *7*, 1–17.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, *17*, 1121–8.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–14.

Reynolds, J. H. & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, *37*(5), 853–63.

Reynolds, J. H. & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–85.

Robinson, D. L. & Petersen, S. E. (1992). The pulvinar and visual salience. *Trends in neurosciences*, *15*(4), 127–32.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., … Vanduffel, W. (2012). Toward a unified theory of visual area V4. *Neuron*, *74*(1), 12–29.

Rolfs, M., Jonikaitis, D., Deubel, H., & Cavanagh, P. (2011). Predictive remapping of attention across eye movements. *Nature Neuroscience*, *14*(2), 252–6.

Rothkopf, C., Ballard, D., & Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 1–20.

Rothkopf, C. a. & Ballard, D. H. (2013). Modular inverse reinforcement learning for visuomotor behavior. *Biological Cybernetics*, *107*(4), 477–90.

Rottschy, C., Eickhoff, S. B., Schleicher, A., Mohlberg, H., Kujovic, M., Zilles, K., & Amunts, K. (2007). Ventral visual cortex in humans: cytoarchitectonic mapping of two extrastriate areas. *Human brain mapping*, *28*(10), 1045–1059.

Saal, H. (2010). Intrinsic Dimensionality of Visual Stimuli at the Center of Gaze. *Publications of the Institute of Cognitive Science*, *12*.

Saalmann, Y. B., Pinsk, M. a., Wang, L., Li, X., & Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science*, *337*(6095), 753–6.

Sapir, a., Soroker, N., Berger, a., & Henik, a. (1999). Inhibition of return in spatial attention: direct evidence for collicular generation. *Nature Neuroscience*, *2*(12), 1053–4.

Sato, T., Murthy, a., Thompson, K. G., & Schall, J. D. (2001). Search efficiency but not response interference affects visual selection in frontal eye field. *Neuron*, *30*(2), 583–91.

Schall, J. & Morel, A. (1995). Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *Journal of Neuroscience*, *15*(6), 4464–4487.

Scheibert, J., Leurent, S., Prevost, A., & Debregeas, G. (2009). The role of fingerprints in the coding of tactile information probed with a biomimetic sensor. *Science*, *323*(March), 1503–1506.

Schlag-Rey, M., Schlag, J., & Dassonville, P. (1992). How the frontal eye field can impose a saccade goal on superior colliculus neurons. *Journal of neurophysiology*, *67*(4), 1003–5.

Schürmann, T. & Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos*, *6*(3), 414–427.

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5), 1–30.

Schütz, A. C., Trommershäuser, J., & Gegenfurtner, K. R. (2012). Dynamic integration of information about salience and value for saccadic eye movements. *Proceedings of the National Academy of Sciences of the United States of America*, 1–6.

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1963). Cultural differences in the perception of geometric illusions. *Science*, *139*, 769–771.

Seo, H. & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, *9*, 1–27.

Serences, J. T., Shomstein, S., Leber, A. B., Golay, X., Egeth, H. E., & Yantis, S. (2005). Coordination of voluntary and stimulus-

driven attentional control in human cortex. *Psychological science*, *16*(2), 114–22.

Serences, J. T. & Yantis, S. (2007). Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cerebral cortex*, *17*(2), 284–293.

Shepherd, S. V., Steckenfinger, S. a., Hasson, U., & Ghazanfar, A. a. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology*, *20*(7), 649–56.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences*, *8*(5), 223–230.

Simoncelli, E. P. & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review Of Neuroscience*, *24*, 1193–1216.

Simons, D. J. & Rensink, R. a. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20.

Simons, D. & Chabris, C. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–74.

Smith, T. J. & Henderson, J. M. (2009). Facilitation of return during scene viewing. *Visual Cognition*, *17*(6-7), 1083–1108.

Smith, T. J. & Henderson, J. M. (2011a). Does oculomotor inhibition of return influence fixation probability during scene search? *Attention, perception and psychophysics*, *73*(8), 2384–2398.

Smith, T. J. & Henderson, J. M. (2011b). Looking back at Waldo : Oculomotor inhibition of return does not prevent return fixations. *Journal of Vision*, *11*, 1–11.

Soltani, A. & Koch, C. (2010). Visual saliency computations: mechanisms, constraints, and the effect of feedback. *Journal of Neuroscience*, *30*(38), 12831–43.

Sprague, N. & Ballard, D. (2003). Eye movements for reward maximization. *Advances in Neural Information Processing Systems*.

Stanton, G. B., Bruce, C. J., & Goldberg, M. E. (1995). Topography of projections to posterior cortical areas from the macaque frontal eye fields. *The Journal of Comparative Neurology*, *353*(2), 291–305.

Steger, J., Wilming, N., Wolfsteller, F., Höning, N., & König, P. (2009). The JAMF Attention Modelling Framework. In *Attention in cognitive systems. lecture notes in computer science.* (pp. 153–165). Berlin Heidelberg: Springer.

Sullivan, B., Johnson, L., Rothkopf, C., Ballard, D., & Hayhoe, M. (2012). The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision*, *12*(13), 1–17.

Sultan, F., Augath, M., Hamodeh, S., Murayama, Y., Oeltermann, A., Rauch, A., & Thier, P. (2012). Unravelling cerebellar pathways with high temporal precision targeting motor and extensive sensory and parietal networks. *Nature communications*, *3*(May), 924.

Swisher, J. D., Halko, M. a., Merabet, L. B., McMains, S. a., & Somers, D. C. (2007). Visual topography of human intraparietal sulcus. *Journal of Neuroscience*, *27*(20), 5326–37.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review Of Neuroscience*, *19*, 109–139.

Tatler, B. W. & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 5.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643–59.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–62.

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *i-Perception*, *1*(1), 7–27.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*, 1–17.

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*, 1–23.

Tatler, B. & Vincent, B. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6), 1029–1054.

Tavakoli, H., Rahtu, E., & Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Image Analysis*, 666–675.

Thomas, N. W. D. & Paré, M. (2007). Temporal processing of saccade targets in parietal cortex area LIP during visual search. *Journal of neurophysiology*, *97*(1), 942–7.

Thompson, K. G., Hanes, D. P., Bichot, N. P., & Schall, J. D. (1996). Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *Journal of neurophysiology*, *76*(6), 4040–55.

Thompson, K. & Bichot, N. (2005). A visual salience map in the primate frontal eye field. *Progress in brain research*, 249–262.

Thompson, K. G., Bichot, N. P., & Sato, T. R. (2005). Frontal eye field activity before visual search errors reveals the integration of bottom-up and top-down salience. *Journal of neurophysiology*, *93*(1), 337–51.

Thompson, K. G., Biscoe, K. L., & Sato, T. R. (2005). Neuronal basis of covert spatial attention in the frontal eye field. *Journal of Neuroscience*, *25*(41), 9479–87.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, *113*(4), 766–86.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

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

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1), 86–94.

Trybula, S. (1958). Some problems of simultaneous minimax estimation. *The Annals of Mathematical Statistics*, *29*(0003-4851), 245–253.

Ungerleider, L. G., Galkin, T. W., Desimone, R., & Gattass, R. (2008). Cortical connections of area V4 in the macaque. *Cerebral cortex*, *18*(3), 477–99.

Ungerleider, L. & Christensen, C. (1979). Pulvinar lesions in monkeys produce abnormal scanning of a complex visual array. *Neuropsychologia*.

Vig, E., Dorr, M., & Cox, D. (2014). Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *Ieee computer vision and pattern recognition (cvpr)*.

Wandell, B. a., Chial, S., & Backus, B. T. (2000). Visualization and measurement of the cortical surface. *Journal of cognitive neuroscience*, *12*(5), 739–52.

Wandell, B. a., Dumoulin, S. O., & Brewer, A. a. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–83.

Wardak, C., Ibos, G., Duhamel, J.-R., & Olivier, E. (2006). Contribution of the monkey frontal eye field to covert visual attention. *Journal of Neuroscience*, *26*(16), 4228–35.

Wardak, C., Olivier, E., & Duhamel, J.-R. (2004). A deficit in covert attention after parietal cortex inactivation in the monkey. *Neuron*, *42*(3), 501–8.

Warnking, J. (2002). fMRI Retinotopic Mapping—Step by Step. *NeuroImage*, *17*(4), 1665–1683.

Williford, T. & Maunsell, J. H. R. J. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology*, 40–54.

Wilming, N., Betz, T., Harst, S., Waterkamp, S., & König, P. (2011). Bayesian modeling of Eye-Movements based on an Analysis of the Conditional Dependence Dtructure between Saccades. In *Osnabrück computational alliance meeting* (Poster).

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and Limits of Models of Fixation Selection. *PLoS ONE*, *6*(9), e24038.

Wilming, N., Harst, S., Schmidt, N., & König, P. (2013). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Computational Biology*, *9*(1), e1002871.

Wilming, N., Jutras, M., Buffalo, E. A., & König, P. (n.d.). Differential contribution of low and high-level image content to eye movements in monkeys and humans. *(In Preparation)*.

Wilming, N., Wolfsteller, F., König, P., Caseiro, R., Xavier, J., & Araújo, H. (2009). Attention Models for Vergence Movements based on the JAMF Framework and the POPEYE Robot. In *Isapp 2009 - proceedings of the fourth international conference on computer vision theory and applications, lisboa, portugal, february 5-8, 2009 - volume 2.* (pp. 429–437). INSTICC Press.

Wright, R. D. & Ward, L. M. (2008). *Orienting of attention*. New York: Oxford University Press.

Wurtz, R. H. & Albano, J. E. (1980). Visual-motor function of the primate superior colliculus. *Annual Review Of Neuroscience*, *3*, 189–226.

Wyss, R., Konig, P., Verschure, P. F. M. J., & König, P. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, *4*(5), e120.

Wyss, R., König, P., & Verschure, P. F. M. J. (2004). Involving the motor system in decision making. *Proceedings. Biological sciences / The Royal Society*, *271 Suppl*, S50–S52.

Yanulevskaya, V., Marsman, J. B., Cornelissen, F., & Geusebroek, J.-M. (2011). An Image Statistics-Based Model for Fixation Prediction. *Cognitive Computation*, *3*(1), 94–104.

Yarbus, A. (1967). *Eye movements and vision* (1st) (B. Haigh & L. Rigss, Eds.). New York: Plenum Press.

Yoshida, M., Itti, L., Berg, D. J., Ikeda, T., Kato, R., Takaura, K., … Isa, T. (2012). Residual attention guidance in blindsight monkeys watching complex natural scenes. *Current Biology*, *22*(15), 1429–34.

Zednik, C. & Jäkel, F. (2014). How does Bayesian reverse-engineering work? In *Proceedings of the 36th annual conference of the cognitive science society*.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological review*, *115*(4), 787–835.

Zénon, A. & Krauzlis, R. J. (2012). Attention deficits without cortical neuronal deficits. *Nature*, *489*(7416), 434–7.

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 32.

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012). Neural activities in v1 create a bottom-up saliency map. *Neuron*, *73*(1), 183–92.

Zhao, Q. & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*, 1–15.

Zhaoping, L. (2011). Neural circuit models for computations in early visual cortex. *Current Opinion in Neurobiology*, *21*(5), 808–15.

Zhaoping, L. & May, K. a. (2007). Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, *3*(4), e62.

Declaration

All experiments reported in this thesis conform with the Declaration of Helsinki and have been approved by the ethics committees of the respective institution (University of Osnabrück, Osnabrück, Germany; Emory University, Atlanta, USA; Wshington University, Seattle, USA). I hereby confirm that I wrote this thesis independently and that I have not made use of resources other than those indicated. I have significantly contributed to all materials used in this thesis. Furthermore, this thesis was neither published in Germany nor abroad, except the parts indicated above, and has not been used to fulfill any other examination requirements.