MODELLING EARLY SPATIAL VISION AND ITS INFLUENCE ON EYE MOVEMENTS IN NATURAL SCENES

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

HEIKO H. SCHÜTT
aus Darmstadt, Deutschland

März, 2018

Tag der mündlichen Prüfung: 24. Juli, 2018

Dekan der Math.-Nat. Fakultät:	Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät:	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Felix A. Wichmann, Dphil.
2. Berichterstatter:	Prof. Dr. Ralf Engbert
3. Berichterstatter:	Prof. Dr. Wolfgang Einhäuser-Treyer

Prüfungskommission:	Prof. Felix A. Wichmann, Dphil
	Prof. Dr. Ralf Engbert
	Prof. Dr. Hanspeter A. Mallot
	Prof. Dr. Philipp Berens

Erklärung

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

"Modelling Early Spatial Vision and its Influence on Eye Movements in Natural Scenes"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, den ....

Datum / Date Unterschrift /Signature

ABSTRACT

As we can only see well with a tiny part of our visual field we need to constantly move our eyes to perceive the world around us. Conversely, our eye movements need to be planned with the information we perceived before. Despite this bidirectional relationship, visual processing and eye movements are typically studied separately. To reunite these fields I design models to predict what we can discriminate and where we look simultaneously. I develop a image-computable spatial vision model, which generalizes classical detection and discrimination data to predict how well arbitrary images can be discriminated. This model fits the classical detection and discrimination data as well as more abstract models, fits natural image masking sensibly and additionally allows me to calculate an experimentally validated internal representation of the stimuli used in eye movement research. Next, I develop statistical methods to evaluate dynamical eye movement models based on direct evaluation of the likelihood of the measured data. These methods are applicable to essentially all eye movement models and provide a solid base for fitting, evaluating and comparing these models. Furthermore, these methods allow Bayesian inference for model parameters and hierarchical models with different parameters for different subjects. Finally, I use the early spatial vision model and the improved evaluation techniques to predict a fixation density from the internal representation generated by the early spatial vision model. Comparing these predictions to other models over time enables me to separate the contributions of bottom-up, top-down, low-level and high-level factors. The combination of my fixation density model with the existing SceneWalk model for the eye movement dynamics results in a mechanistically plausible model which predicts both eye movement and discrimination experiments. Building on the foundations I made, future research might extend my model to include higher level processing, to include more dependencies within scanpaths and to include a peripheral decline in visual processing to further expand our understanding of eye movements and visual perception.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, 124(4), 505–524.

Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, 17(12), 12:1-35.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2016). Influence of initial fixation position in scene viewing. *Vision Research*, 129, 33–49.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, 17(13), 3:1-18.

This thesis is accompanied by a Statement what my exact contributions were to these papers and to the following two so far unpublished manuscripts:

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R. & Wichmann, F. A. (Manuscript). Disentangling top-down vs. bottom-up and low-level vs. high-level influences on eye movements over time.

Rothkegel, L. O. M., Schütt, H. H., Trukenbrod, H. A., Wichmann, F. A., & Engbert, R. (2018). Searchers adjust their eye movement dynamics to the target characteristics in natural scenes. arXiv:1802.04069 [q-bio].

STATEMENT OF CONTRIBUTIONS ACCORDING TO §9(2):

My Doctoral Thesis was part of a joint DFG-Funded project of my main supervisor Prof. Felix A. Wichmann and Prof. Ralf Engbert in Potsdam. Major contributions on the publications written in this project were made additionally by Lars O.M. Rothkegel, another PhD-student working on the project in Potsdam and Dr. Hans A. Trukenbrod who always accompanied our project from Potsdam. Furthermore Prof. Sebastian Reich, a mathematics professor from Potsdam helped with input on mathematical details of one paper.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, 124(4), 505–524.

This article is largely my own work. I originally conceived the idea that one can directly calculate a likelihood for the kind of models we analyse usually and performed all analyses presented in the paper. Lars O. M. Rothkegel and Hans A. Trukenbrod provided the data I used for the evaluation and advised me about eye movements. Sebastian Reich introduced me to some more advanced data assimilation techniques and checked the mathematical integrity of my approach. Finally, Felix A. Wichmann and Ralf Engbert supervised me and provided the necessary work environment. All authors also contributed to revisions of the paper.

Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, 17(12), 12:1-35.

This article is also largely my own work. I designed and implemented the model and performed all analyses. Felix A. Wichmann provided the data from earlier experiments and also supervised and advised me about early visual processing.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2016). Influence of initial fixation position in scene viewing. *Vision Research*, 129, 33–49.

The experiments described in this paper were performed by Lars O. M. Rothkegel and he also led the writing of the article. My contributions to this paper were largely advisory regarding the choice of stimuli and regarding the interpretation and presentation of the data.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, 17(13), 3:1-18.

Again, the experiments in this paper were performed by Lars O. M. Rothkegel and my contributions were largely advisory. In this case I suggested the explanation that only the duration since image onset matters for the strength of the central fixation bias. Otherwise I mainly contributed to presentation and write up of this article.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R. & Wichmann, F. A. (Manuscript). Disentangling top-down vs. bottom-up and low-level vs. high-level influences on eye movements over time.

This manuscript was led by me. I performed all analyses presented in this manuscript and wrote the first draft of this manuscript. However, Lars O. M. Rothkegel contributed more to this manuscript than to the earlier papers as he again collected all eye movement data necessary for these analyses. These datasets are exceptionally large and the corpus dataset was collected especially for these analyses, which accentuates his contribution to this manuscript. As for all other papers, Hans A. Trukenbrod, Felix A. Wichmann and Ralf Engbert provided advise and supervision and all authors contributed to revisions of the manuscript.

Rothkegel, L. O. M., Schütt, H. H., Trukenbrod, H. A., Wichmann, F. A. & Engbert, R. (Manuscript). Searchers adjust their eye movement dynamics to the target.

The experiment described in this paper was led by Lars O. M. Rothkegel again. This time however, I directly contributed to the design of the study and provided means to generate the stimuli. Furthermore I contributed saliency based and image patch based analyses and wrote (small) parts of the paper originally. As for all other papers, Hans A. Trukenbrod, Felix A. Wichmann and Ralf Engbert provided advise and supervision and all authors contributed to revisions of the manuscript.

Tübingen, März 2018,

Lars O.M. Rothkegel	Hans A. Trukenbrod	Ralf Engbert
Felix A. Wichmann	Sebastian Reich	Heiko H. Schütt

ACKNOWLEDGEMENTS

My thanks go to my supervisors Felix Wichmann and Ralf Engbert whose wise supervision made this thesis possible. Additionally I want to thank my companions in this project Lars Rothkegel and Hans Trukenbrod. I can not think of a better team.

Also, I want to thank everyone from Tübingen & Potsdam who enjoyed lunches, conferences and life in general with me during my thesis.

I am grateful to share my life with my Anna, who made all the time outside the university so much better.

Finally, my thanks go to my parents for raising me to become the one I am today and to my family and friends for accompanying and guiding me through my life.

CONTENTS

1 INTRODUCTION 1
1.1 Why models? 2
1.2 The model framework 3
1.2.1 Human behaviour as description level 3
1.2.2 Natural image stimuli as experimental approach 4
1.3 Early visual processing 5
1.3.1 Earlier models with similar aims 5
1.3.2 Advantages of image computability 6
1.4 Eye movement models 6
1.4.1 Restriction to static natural scenes 6
1.4.2 Restriction to modelling fixation locations 7
1.5 Classes of eye movement models 8
1.5.1 Domain specific models 8
1.5.2 Top-down, bottom-up, low- and high-level 9
1.5.3 Attention: low-level, top-down 9
1.5.4 Classical saliency: low-level, bottom-up 9
1.5.5 Tasks and intentions: high-level, top-down 10
1.5.6 Modern saliency models: high-level, bottom-up 10
1.5.7 Optimal control 10
1.5.8 Systematic tendencies 11
1.5.9 Difficulty of evaluation 11
1.5.10 The SceneWalk model 12
1.6 Connecting early vision and eye movements 12
1.7 Data source 12
1.8 Outline 13
2 AN IMAGE COMPUTABLE EARLY VISION MODEL 15
2.1 Introduction 16
2.1.1 History and classical experiments in spatial vision 17
2.1.2 Outline 18
2.2 Model Description 18
2.2.1 Preprocessing 20
2.2.2 Decomposition 21
2.2.3 Normalization & nonlinearity 24
2.2.4 Noise and decoding 27
2.2.5 Calculating thresholds 29
2.2.6 Parameter fits 29
2.2.7 Data for model evaluation 30
2.3 Results 30
2.3.1 Classical psychopysical results 30
2.3.2 Natural scene masking database 35
2.3.3 Different parameter sets 37
2.3.4 Analysis of the models representation 38
2.3.5 Sparseness 39
2.3.6 Optimized stimuli 40
2.4 Discussion 42

2.4.1 Comparison to earlier models . . . 43
2.4.2 Potentially controversial details . . . 44
2.4.3 Limitations of the presented model . . . 47
2.4.4 Applications in and beyond spatial vision . . . 48
3 LIKELIHOOD-BASED EVALUATION OF DYNAMICAL COGNITIVE MODELS 49
3.1 Introduction . . . 50
3.2 Likelihood computation for dynamical models . . . 52
3.2.1 The likelihood for dynamical models based on discrete observations 53
3.2.2 Model details . . . 55
3.2.3 Competing models . . . 57
3.3 Estimation of model parameters . . . 58
3.3.1 Maximum likelihood estimation . . . 59
3.3.2 Bayesian inference . . . 60
3.3.3 Results on model parameter estimation . . . 61
3.3.4 Inter-Subject differences and hierarchical models . . . 65
3.4 Model comparison in the likelihood approach . . . 67
3.4.1 Results on model comparison . . . 70
3.5 Goodness-of-fit for specific measures and spatial statistics . . . 72
3.6 Discussion . . . 75
3.7 Conclusion . . . 79
4 CONNECTING EARLY VISION AND EYE MOVEMENTS 81
4.1 Introduction . . . 82
4.2 Methods . . . 84
4.2.1 Stimulus presentation . . . 84
4.2.2 Measurement of eye movements . . . 84
4.2.3 Corpus dataset . . . 85
4.2.4 Natural image search . . . 87
4.2.5 Analysing fixation locations . . . 88
4.2.6 Gold standard analyses . . . 88
4.2.7 Comparing fixation densities . . . 88
4.2.8 Evaluation of saliency models . . . 89
4.2.9 Tested saliency models . . . 90
4.3 Results: Corpus . . . 91
4.3.1 Overall saliency model performance . . . 91
4.3.2 Temporal aspects . . . 93
4.3.3 Saliency models over time . . . 94
4.3.4 Density of the first fixation . . . 95
4.4 Results: Visual Search . . . 97
4.4.1 Fixation densities . . . 97
4.4.2 Saliency models . . . 98
4.4.3 Fixated patches . . . 99
4.4.4 Predicting search performance . . . 100
4.5 Discussion . . . 101
4.5.1 Bottom-up vs. top-down . . . 102
4.5.2 Low-level vs. high-level . . . 103
4.5.3 Physiological substrate . . . 104
4.5.4 Future prospects . . . 104
4.5.5 Conclusion . . . 105
5 DISCUSSION 107

5.1 Embedding of results . . . 108
5.1.1 Early vision model . . . 108
5.1.2 Evaluation methods for eye movement models . . . 108
5.1.3 Connecting early visual processing to eye movements . . . 109
5.1.4 Other studies I contributed to . . . 109
5.2 Alternative approaches . . . 111
5.2.1 Other evaluation techniques . . . 111
5.2.2 Non-mechanistic models . . . 113
5.2.3 Non-image-computable approaches . . . 114
5.2.4 Search models . . . 115
5.2.5 Fixation durations . . . 115
5.2.6 Attention models . . . 116
5.3 Controversies . . . 117
5.3.1 Automated vs. cognitive control . . . 117
5.3.2 Inhibition of return . . . 118
5.3.3 Maps vs. Objects . . . 119
5.4 Directions for future research . . . 120
5.4.1 Peripheral processing . . . 120
5.4.2 Higher-level processing . . . 121
5.4.3 Crowding: Peripheral restriction on higher levels . . . 121
5.4.4 Eye movement dynamics . . . 123
5.4.5 Dynamics and image content . . . 124
5.4.6 Statistical improvements . . . 124
5.5 Conclusion . . . 125
A MATHEMATICAL DETAILS 127
A.1 Fitting . . . 128
A.2 Derivatives of the Model . . . 132
A.2.1 Likelihood from signal to noise ratio . . . 133
A.2.2 Decoding . . . 133
A.2.3 Normalization . . . 134
A.2.4 Decomposition . . . 135
A.2.5 Preprocessing . . . 136
A.3 Optimizing stimuli . . . 136

1

INTRODUCTION

Schreiben ist gut, Denken ist besser.
Klugheit ist gut, Geduld ist besser.

Siddhartha, Hesse (1922)

Visual perception is our primary sense to perceive the world surrounding us and has thus always fascinated human thinkers. From Buddha and Konfuzius to the ancient Greek philosophers to the arabic philosophy in the middle ages and the philosophers of the illumination seeing was always considered a metaphor for understanding and believing.

At the same time we know all too well, that our visual perception is limited. Our visual system is sensitive only to a tiny part of electromagnetic radiation and we can only see sharply with our fovea, a small area in our retina, which contains the highest receptor density. Even of this limited supply of information we do not perceive all, but may miss things we do not attend to. Thus, we have to constantly shift our eyes and attention to gather the information we require.

As eye movements and attention shifts are integral parts of human visual perception and possibly the narrowest bottleneck on the path to perception, their analysis may help us to understand perception. Furthermore eye movements might inform us which information is interesting or relevant to observers hinting at their goals, thoughts and desires.

However, understanding what drives eye movements requires understanding what information can be perceived when some eye movements are used. Conversely, understanding perception in natural circumstances requires an understanding of eye movements. Thus, these two topics are intimately related, although they have been largely studied separately. Most studies on what can be perceived (especially peripherally) enforce fixation, i.e. the suppression of eye movements. At the same time most studies on eye movements and attention ignore our understanding of visual perception and its limitations in both the design of studies and the design of models to explain their results.

In this thesis I try to bridge some of the gap between these two fields and hopefully move the models for both parts slightly forward. To do so, I aimed to build a model which predicts eye movements in natural scenes, but takes the limitations of early visual processing into account, i.e. uses only information humans can perceive and treats perceptually similar images similarly.

1.1 WHY MODELS?

Declaring a model as the main aim for my thesis makes the assumption that such models can advance the scientific process. This assumption is sometimes questioned by more experimentally inclined scientists. To counter this objection, I present some opportunities here, how formal models can contribute to scientific progress in general.

The first advantage of formal models or theories is that they can provide quantitative predictions for experiments and phenomena, while non-formal theories are restricted to qualitative predictions. This allows many additional strong tests of the underlying theories, especially in cases where multiple theories predict the same direction of an effect.

Additionally, quantitative predictions are valuable in themselves. For many applications of our understanding it is important to quantify the effects. Especially, rationally choosing from multiple possible actions requires a quantitative prediction of the expected results (Berger, 2013). To be applicable to real world decisions, further restrictions apply: The models additionally need to be predictive outside the lab and the available information in real world applications might be restricted. Nonetheless, making quantitative predictions is a necessary condition to apply any rational decision mechanisms.

Formal models are especially helpful to understand complex situations, systems or behaviour. In complex situations many experimental measures depend on each other such that even communicating a proposed explanation of the situation may require considerable formalization of the ideas. Also an explanation of complex behaviour will usually require

the combination of many studies and experimental results, which is facilitated by the design of a formal model. Furthermore, manipulating a formal model by removing or adding parts allows one to test the effects of model parts on distant predictions. These predictions can then be tested to understand which theoretical concepts are important to explain which connections.

Finally, formal models predict the whole behaviour which might contain dependencies, which emerge from the interplay of different parts and were not explicitly incorporated into the model. Such emerging behaviours are an important source for theoretical advances, providing new predictions from theory and thus ways to test the theory. If such emerging predictions are confirmed by new data, these data provide perhaps the strongest evidence for a model one could hope for. In contrast, without formalizing the ideas, the chances that a model surprises its inventor are slim, because any prediction has to be deducted from the theory by the inventor themself.

1.2 THE MODEL FRAMEWORK

The two most import dimensions models of human behaviour vary over are: The naturalism of the modelled experimental conditions, which ranges from perfectly controlled laboratory experiments to free behaviour. And the level of explanation which varies from the behaviour of single molecules to the full behaviour. This second dimension is also strongly related to the choice of model species as full human behaviour can only be studied in humans while more and more distantly related species are preferred for studying increasingly lower levels of explanation.

1.2.1 *Human behaviour as description level*

The level of description used in a model is mostly determined by the aims of a model, because a model needs to make predictions at the same level as the research questions are formulated. Therefore, all modelling in this thesis is described at the level of human behaviour, because I aimed to understand human behaviour and only evaluate my models on human behaviour. I will glance over connections to studies in other animals and physiology in general, although many parts of my models can be mapped to anatomical structures which seem to perform functions similar to the model parts. Namely the early visual processing shall correspond to the processing performed by the cascade from the retina through the lateral geniculate nucleus up to primary visual cortex and many aspects of the early vision model are found in models of neurons in early visual cortex as well (e.g. Cavanaugh, Bair, & Movshon, 2002a; Heeger, 1992; Heeger, Simoncelli, & Movshon, 1996). Similarly, central parts of our eye movement control model seem to be encoded by the superior colliculus, which seems to encode a saliency map (in its superficial layers) and the final priority map for the choice of fixation location (in its intermediate layers White, Berg, et al., 2017; White, Kan, Levy, Itti, & Munoz, 2017; White & Munoz, 2011). Alternative physiological substrates could be the frontal eye field (Johnston & Everling, 2011) and the posterior parietal cortex (Paré & Dorris, 2011), which contain similar priority maps and contribute to eye movement control, but are usually associated with higher level control not with bottom up saliency. Thus, the models I describe do not contradict our knowledge about physiological processes, but none of the models and experiments I present is concerned with physiology directly.

1.2.2 *Natural image stimuli as experimental approach*

The naturalness of the experimental conditions heavily influences what kind of data is available for evaluation. Simple, controlled stimuli have the advantage that they can be tuned to the exact research questions by isolating individual variables of interest (Rust & Movshon, 2005). Exploiting this adaptability, many researchers use simple stimuli for research on both early visual processing (e.g. Campbell & Robson, 1968; Goris, Zaenen, & Wagemans, 2008; Henning, Hertz, & Broadbent, 1975; Legge & Foley, 1980; Meese, Georgeson, & Baker, 2006; Meese & Holmes, 2002; Nachmias & Sansbury, 1974) and eye movements (e.g. Aagten-Murphy & Bays, 2017; Hallett, 1978; Schütz, Trommershäuser, & Gegenfurtner, 2012; Theeuwes, Kramer, Hahn, & Irwin, 1998). Furthermore, there are models available for these experiments, which claim to explain many of these data (e.g. Campbell & Robson, 1968; Foley & Legge, 1981; Goris, Putzeys, Wagemans, & Wichmann, 2013; Wolfe, 1994). However, different simple stimuli are usually used for early visual processing studies (gratings mostly) and for eye movement research (coloured geometric figures, isolated objects and flashes) and different parameters of the stimuli are manipulated, which makes generalizations from one field to the other hard. Additionally, models for simple stimuli and experiments are often hard to generalize to other situations.

More complex, natural stimuli on the other hand have the advantage, that they include more attributes and regularities, which exist in the natural environment and which might be important to understand eye movements (context and scene understanding for example: Cornelissen & Võ, 2017; Torralba, Oliva, Castelhano, & Henderson, 2006) or early visual processing (efficient coding for example: H. Barlow, 2001; H. B. Barlow, 1969; Ganguli & Simoncelli, 2014; Olshausen & Field, 1996). Control of such naturalistic experiments is limited though, which is problematic for modelling, because the uncontrolled variations between experimental runs are near impossible to include into the model and thus add unexplained variance. Real world stimuli which are not presented on a screen (Land, Mennie, & Rusted, 1999) are not immediately available in digital form and thus add another level of complexity how the stimuli should be represented and which aspects of this representation should be part of the input to the model. Finally, naturalistic stimuli and experiments reintroduce the complexity that early vision and eye movement experiments differ in the stimuli and experimental paradigms used.

As a compromise on this continuum I chose the behaviour when exploring static natural images on a screen. Natural images are popular stimuli for experiments on both early visual processing (e.g. Alam, Vilankar, Field, & Chandler, 2014; J. Freeman & Simoncelli, 2011; Wallis & Bex, 2012) and eye movements (e.g. Açık, Onat, Schumann, Einhäuser, & König, 2009; Bylinskii et al., 2016; Cajar, Engbert, & Laubrock, 2016; Cajar, Schneeweiß, Engbert, & Laubrock, 2016; Einhäuser & König, 2003; Einhäuser & Nuthmann, 2016; Einhäuser, Rutishauser, & Koch, 2008; Einhäuser, Spain, & Perona, 2008; Neider & Zelinsky, 2006). Natural images contain many regularities present in our environment, although some additional regularities and biases are introduced by photographers to make aesthetically pleasing images (Cooper, Piazza, & Banks, 2012; Wichmann, Drewes, Rosas, & Gegenfurtner, 2010). Also participants are familiar with viewing images, the images can be presented easily on computer screens for experiments, allow exact replications of the stimulus configuration and procedure, and can be manipulated in high detail by using image processing. Additionally, basic image-computable models (Khaligh-Razavi & Kriegeskorte, 2014; D. L. Yamins & DiCarlo, 2016) were available of both early visual processing (e.g. Teo & Heeger, 1994; Watson & Solomon, 1997) and eye movement behaviour (Judd, Ehinger, Durand, & Torralba, 2009; Kienzle, Franz, Schölkopf, & Wichmann, 2009;

Zelinsky, 2008) before I started my thesis, fuelling hopes that a combined model using natural images as input could be successful.

Combined with the notion that early visual processing should provide the knowledge base for the control of eye movements, these specifications already fix the rough form of my model: I develop a model of early visual processing to arbitrary images, which results in an internal representation of the image. This internal representation serves as the basis for a saliency map, which signifies the regions of the image which are interesting for our model observer. Finally, a model of the scanpath dynamics describes how observers explore the areas given by this map to create full scanpaths.

1.3 EARLY VISUAL PROCESSING

The first inevitable step to realize my model is the design of an image-computable model (Khaligh-Razavi & Kriegeskorte, 2014; D. L. Yamins & DiCarlo, 2016) of early spatial visual processing. The model must calculate an internal representation of arbitrary images to make predictions for the image stimuli used in the eye movement experiments.

This first restriction of image-computability already excludes a broad range of models of early spatial vision, which directly operate on the parameters of simple stimuli like spatial frequency or contrast of gratings (e.g. Foley, 1994; Georgeson, Wallis, Meese, & Baker, 2016; Goris et al., 2013; Itti, Koch, & Braun, 2000). These models and the corresponding experimental results which motivated them (e.g. Campbell & Robson, 1968; Goris et al., 2008; Legge & Foley, 1980; Meese et al., 2006; Meese & Holmes, 2002) nonetheless shape our understanding of the first steps of visual processing. This understanding fortunately provides a unified account of the behavioural and physiological experimental results, such that the general structure of the model is not controversial.

1.3.1 *Earlier models with similar aims*

It has been tried to apply early vision models to images (e.g. Bradley, Abrams, & Geisler, 2014; Teo & Heeger, 1994; Watson & Solomon, 1997). However, the used models were either extremely simplified, concentrating on carefully tuned spatial frequencies, orientations and experiments (Teo & Heeger, 1994; Watson & Solomon, 1997) or the processing in the model is tuned to the target to be detected (Bradley et al., 2014; Itti et al., 2000). Thus, the previous models of early spatial vision were not usable for my purposes, because they either handled only specific experiments and stimulus ranges or do not make their internal representation explicit.

Another source for an image computable early spatial vision model are models predicting fixation locations which traditionally relied on representations inspired by V1 (Itti & Koch, 2000; Judd et al., 2009). However, none of these models was evaluated on detection or discrimination tasks explicitly. Instead these models were only inspired by the models of early visual processing, which is only a weak test of the hypothesis that early visual processing feeds into perception and into the programming of eye movements.

A third source for image computable models of early visual processing are image quality metrics (Laparra, Ballé, Berardino, & Simoncelli, 2016; Wang, Simoncelli, & Bovik, 2003). Models of image quality superficially solve the same task of predicting which differences between images are visible and relevant to humans. However, these models do not necessarily aim to process stimuli the same way as the human visual system does. Instead such metrics aim to provide computationally "cheap" solutions, which fit human ratings

reasonably well. Additionally these models aim to predict not only which errors are visible, but also how disturbing the visible errors are (Wang, Bovik, Sheikh, & Simoncelli, 2004), which means they may not only represent early visual processing. Furthermore, these models do not necessarily compute any internal representation one could use for modelling further processing.

1.3.2 *Advantages of image computability*

As no suitable image-computable model of early visual processing existed, the first step of my thesis was to implement such a model. This model is described in detail in our publication on the topic (Schütt & Wichmann, 2017) and the corresponding Chapter 2 of this thesis.

Beyond the need to develop an image-computable early spatial vision model to investigate its interactions with eye movements, an image-computable early spatial vision model is an achievement in its own right. For example, image-computable models of early vision allow more thorough tests of the model, especially towards more natural stimuli. As first steps in this direction, I present an evaluation on natural image stimuli, fixated image locations and some optimized example stimuli, which give more insight into the processing of the model in Chapter 2. Beyond these analyses there are specialized methods to compare models like maximum differentiating stimuli (Wang & Simoncelli, 2008). Furthermore, an image computable model can be evaluated on existing datasets, which measure the visibility of various image distortions (e.g. Alam et al., 2014; Ponomarenko et al., 2008). Such evaluations allow efficient model comparisons and might highlight errors of the model, which are not apparent with the simplified artificial stimuli classically employed in psychophysics. This use of natural stimuli is even accepted by prominent proponents of the use of artificial stimuli in vision research (Rust & Movshon, 2005).

1.4 EYE MOVEMENT MODELS

Turning to my final modelling goal of predicting natural eye movements, the first thing to point out is that eye movements are a broad field in their own right. There is an overwhelming breadth of scientific studies on eye movements ranging from the control of single types of eye movements (saccades, smooth pursuit & fixational eye movements) in extremely controlled laboratory tasks (e.g. Aagten-Murphy & Bays, 2017; Engbert & Kliegl, 2003; Hallett, 1978; Schütz et al., 2012) up to free eye movement behaviour during natural tasks (e.g. Land & Lee, 1994; Land et al., 1999; 't Hart & Einhäuser, 2012). Also the models range from specific models for circumscribed domains like reading (e.g. Engbert, Nuthmann, Richter, & Kliegl, 2005; Reichle, Pollatsek, Fisher, & Rayner, 1998) to general models which aim to describe eye movement behaviour in general contexts (Adeli, Vitu, & Zelinsky, 2017; Engbert, Trukenbrod, Barthelme, & Wichmann, 2015; Itti & Koch, 2000; Treisman & Gelade, 1980; Trukenbrod & Engbert, 2014; Tsotsos et al., 1995).

1.4.1 *Restriction to static natural scenes*

To reduce this broad scope to a manageable regime, I focus on eye movements in static photographs of natural scenes shown on a display system. This restriction still allows the inclusion of a task, relatively natural stimulation, the complex dynamics of eye movement

behaviour and the natural dependencies between eye movements (Tatler & Vincent, 2008, 2009), but removes some experimental, theoretical and computational problems. For experiments this restriction is helpful as images on displays can be controlled accurately, such that variability in visual stimulation is not problematic and the stimuli can be chosen and adjusted easily to adjust the material to the research questions or task difficulty (as we did for our search data: Rothkegel, Schütt, Trukenbrod, Wichmann, & Engbert, 2018). Furthermore, the data quality of eye movement recordings from a head stabilized laboratory environment is still considerably better than the one of data recorded with mobile eye tracking glasses (Engbert, Rothkegel, Backhaus, & Trukenbrod, 2016). As a theoretical advantage we are able to exclude some influences which would have to be considered with even more natural stimuli like motion in the stimulus—which requires the consideration of smooth pursuit—the coupling of eye movements with other actions of the participant (Land et al., 1999)—which requires the tracking and/or control of these actions and depth—which requires vergence and accommodation eye movements. The visual processing of stimulus attributes not contained in a static image like motion, disparity or defocus is still actively researched (Georgeson et al., 2016; Meese et al., 2006) and most importantly in this context not contained in my model of early visual processing. Finally the restriction to static images on a display significantly reduces the computational load during the computation of our models, because the image processing—which is computationally most expensive—is required only once while any changing stimuli would require image processing over time. This additional load would limit the complexity of the models further, because the models I propose here already use our considerable computational resources to full capacity.

1.4.2 *Restriction to modelling fixation locations*

Even in static images eye movements are a complex behaviour. Many aspects of scanpaths can be interesting for researchers. For example researchers are interested in: fixation locations (Einhäuser, Rutishauser, & Koch, 2008; Einhäuser, Spain, & Perona, 2008; Itti & Koch, 2001; Kienzle et al., 2009), saccade lengths (Ramos Gameiro, Kaspar, König, Nordholt, & König, 2017; Tatler, Baddeley, & Vincent, 2006), other saccade properties (Ludwig & Gilchrist, 2002; Smit & van Gisbergen, 1990), fixation durations (Hooge & Erkelens, 1998), fixational eye movements (Engbert & Kliegl, 2003) and all dependencies between these measures between each other, between consecutive eye movements (Tatler & Vincent, 2008, 2009; Wilming, Harst, Schmidt, & König, 2013) and/or over time (Over, Hooge, Vlaskamp, & Erkelens, 2007; Tatler, 2007). Furthermore, many factors influence eye movements including low level (Itti & Koch, 2001; Kienzle et al., 2009) and high level (Stoll, Thrun, Nuthmann, & Einhäuser, 2015; Torralba et al., 2006) image properties, task and experimental situation (Castelhano, Mack, & Henderson, 2009; Greene, Liu, & Wolfe, 2012; Nuthmann, 2017; Yarbus, 1967), systematic tendencies (Tatler & Vincent, 2008, 2009) and individual differences (Castelhano & Henderson, 2008; Hayes & Henderson, 2017). Finally, eye movements also change the input to the visual system, such that eye movements are always an interaction with the environment even if the world the observer explores is itself stationary.

This broad range of relationships and complications highlights the need for formalized models to link the experimentally observed effects and to unify them into an integrated theory. At the same time the breadth of the field and the observations lead to a broad range of models with diverse aims, theoretical justifications and internal mechanisms, in contrast to the situation for early spatial visual processing.

In this thesis I will focus on modelling fixation locations. Fixation durations have been modelled with substantial sophistication as well (e.g. R. Carpenter, 1999; Nuthmann, Smith, Engbert, & Henderson, 2010; Tatler, Brockmole, & Carpenter, 2017; Trukenbrod & Engbert, 2014). These models are typically based on diffusion models to produce the skewed shape of the fixation duration distribution. Main controversies include the questions whether there is one timer (Trukenbrod & Engbert, 2014) or a race between different target locations (Tatler et al., 2017), up to which point in time visual input can influence the timing of a saccade and what stimulus, previous eye movement and task properties influence the fixation duration (Nuthmann, 2017; Nuthmann et al., 2010). The last question raises the follow up question for each influence factor, whether it influences the fixation duration immediately or adjusts the timing of future fixations (Trukenbrod & Engbert, 2014).

If fixation durations were independent of the chosen fixation locations, one could ignore the modelling of fixation durations for the modelling of fixations locations. However, there is accumulating evidence that this is not the case (Einhäuser & Nuthmann, 2016; Tatler et al., 2017). Especially regions which are fixated more are also fixated longer on average (Einhäuser & Nuthmann, 2016), but the fixations before likely saccades (towards frequently fixated regions) are shorter (Tatler et al., 2017). Furthermore, dependencies between eye movements, which I describe in more detail below, frequently affect both the likelihood of saccades to specific locations and the fixation duration before the saccade (Over et al., 2007; Smith & Henderson, 2009; Tatler & Vincent, 2008, 2009). Thus, it seems likely that the choices when and where to a saccade is executed are coupled and may be made by a common mechanism. In this thesis I will nonetheless restrict myself to the location aspect of eye movement control to make solutions feasible at this point in time. In the future these relationships might help to understand and model the dynamical aspects and dependencies over scanpaths in more detail.

1.5 CLASSES OF EYE MOVEMENT MODELS

Caused by the breadth of the field of eye movement research many different approaches exist to build models of them, which have not yet converged to a standard model. Thus, I discuss these different classes of models individually below.

1.5.1 *Domain specific models*

The first model class to mention are domain specific models. Such models are concerned with eye movements in specific tasks or situations. Such specific models exist with especially high sophistication for reading (Engbert et al., 2005; Reichle et al., 1998; Reichle, Rayner, & Pollatsek, 2003). These models rely heavily on the sequential nature of text, its separation in words and letters, and linguistic properties of words, which have no equivalent for eye movements in other contexts. Thus, the bearing of these eye movement models in natural scenes is small. Less formalized ideas also exist for eye movements while driving (Chapman, Underwood, & Roberts, 2002; Land & Lee, 1994; Underwood, 2007; Underwood, Chapman, Brocklehurst, Underwood, & Crundall, 2003), while viewing faces (Haxby, Hoffman, & Gobbini, 2002; Peterson & Eckstein, 2012, 2013) or while playing chess (Reingold & Sheridan, 2011). Neither of these models provide guidance for the development of models for eye movements in natural scenes. One conclusion one might take

from these studies is that eye movements can be adjusted and trained to specific tasks and indeed change with experience in the specific task (Reingold & Sheridan, 2011).

1.5.2 *Top-down, bottom-up, low- and high-level*

The more general theories and models about eye movement control in less constraint tasks like scene viewing can be divided along two dimensions: Top-down vs. bottom-up control, i.e. the question whether eye movements are determined primarily by the subject or by its environment, and low- vs. high-level accounts, i.e. whether the factors governing eye movements are computed from simple low-level features as computed early in the visual hierarchy or from high-level features like objects and scene configuration. These two dimensions can vary independently and I present examples for all combinations below, although classically the two competing theories are saliency—a low-level, bottom-up account—and cognitive control—a high-level, top-down account.

1.5.3 *Attention: low-level, top-down*

In the attention literature eye movements are usually regarded as overt attention shifts, which are modelled in the same form as covert attention shifts (Tsotsos et al., 1995). As attention is largely conceived as a top-down process, this implies a cognitive, top-down control of attention and of eye movements. However, attention effects are typically modelled to reweigh different low-level features like colours, orientations or spatial locations, especially when modelling visual search (Müller & Krummenacher, 2006; Wolfe, Cave, & Franzel, 1989; Wolfe & Horowitz, 2004). This connection to low-level features is especially prominent in the Feature Integration Theory (Treisman & Gelade, 1980) and all its descendants, as this theory proposes that attention needs to be directed to an object to combine the low-level features into a high-level description in the first place.

One important aspect of attention models, which is largely ignored by other eye movement models, is that attention should change visual processing (Tsotsos et al., 1995), changing the input to the processes which ultimately guide eye movements. Furthermore, attention is coupled to eye movements (Deubel & Schneider, 1996) and this coupling can also be seen in natural scene viewing (Cajar, Schneeweiß, et al., 2016). These attentional effects are interesting research aims in their own right. However, I left out the influence of attention on visual processing as it seemed out of scope for this thesis. In the future the attention effects caused by eye movements might explain some of the sequential dependencies between fixations.

1.5.4 *Classical saliency: low-level, bottom-up*

One descendent of Feature Integration Theory further formalized the idea of low-level features governing eye movements such that conspicuous objects on these features attract attention via a central saliency map (Koch & Ullman, 1985). Linking these ideas to eye movements, Itti, Koch, and Niebur (1998) formed the first image-computable model of human eye movement control, dropping the last top-down influences on the way.

These saliency models, which assume that some visually salient image properties attract fixations largely independent of other influences (Itti & Koch, 2001), are the classical approach to use early vision ideas to predict eye movements and many incarnations of this type of model followed (e.g. Harel, Koch, & Perona, 2006; Itti & Koch, 2000; Itti

et al., 1998; Kienzle et al., 2009). Further strengthening the connection of these original models to low-level features, they were linked to models of early visual processing (Itti et al., 2000) and to the physiology of primary visual cortex (Z. Li, 2002). Although these models provide some predictive value for fixation locations, they never predicted fixations particularly well (Bylinskii et al., 2016).

1.5.5 *Tasks and intentions: high-level, top-down*

The classical competing theory to saliency is the idea that eye movement control is a high level cognitive process. This theory claims that eye movement control is based on expectations about the configuration of scenes (Cornelissen & Võ, 2017; Henderson, Weeks Jr, & Hollingworth, 1999; Mohr et al., 2016) and scene gist (Torralba et al., 2006) or on task instructions (Einhäuser, Rutishauser, & Koch, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007; Yarbus, 1967). These ideas are supported by experimental data, which show differences caused by changes in these influence factors. However, formalized models of these processes are rare and strongly simplified (Torralba et al., 2006). This lack of formal models is understandable, as the high level concepts have not been formalized sufficiently themselves to base complex models on them, although there are attempts to extract higher level image properties like image category (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017), object locations (Lin et al., 2014), or image clutter (Yu, Samaras, & Zelinsky, 2014). As these extraction methods further improve, future eye movement models might be able to include these information.

1.5.6 *Modern saliency models: high-level, bottom-up*

A different critique of low-level, bottom-up models attacked only the low-level aspect, claiming that objects and their properties guide eye movements instead of low-level features. This view lead to models for the prediction of fixation locations based on the location of objects and their parts (Einhäuser, Spain, & Perona, 2008; Stoll et al., 2015), which at their time consistently beat the saliency models in prediction quality, but were not image-computable, i.e. required manual marking of the objects in the image.

As it was noted in computer vision that objects or more complex features may predict fixation locations better (Einhäuser, Spain, & Perona, 2008; Judd et al., 2009), the theoretical justification of saliency as a direct reflexive influence of the low-level features was largely abandoned in this community. As a consequence, more modern saliency models use highly complex features, which encode where salient objects like text or faces are located (Borji & Itti, 2013; Judd et al., 2009; Kümmerer, Wallis, & Bethge, 2015), which represents a theoretical shift towards high-level image properties. Most recently the predictive power of these models based on high-level image features could be improved substantially by using features from deep neural networks trained on object recognition (e.g. Kümmerer et al., 2015). This approach is the most successful approach to predict fixation locations to date.

1.5.7 *Optimal control*

The most formalized high level control theory is optimal control. This idea of an optimal observer model had been successful for some perceptual processes, especially for cue combination (Ernst & Banks, 2002; Geisler, 2011; Landy, Maloney, Johnston, & Young,

1995), although observers typically show some deviations from optimality (Adler & Ma, 2017; Rosas, Wagemans, Ernst, & Wichmann, 2005). This optimal observer theory could be extended to visual search (Najemnik & Geisler, 2005, 2008), given a homogeneous background and measures of the visual decline into the periphery. Furthermore, a simplified implementation of ideal search was proposed (Najemnik & Geisler, 2009), addressing the common critique that the true optimal observer is too complex to be computed by humans for any natural environment. There are some other situations however, where humans do not choose their eye movements optimally (Nowakowska, Clarke, & Hunt, 2017). Thus, optimal behaviour seems not to be a full description of human behaviour. Nonetheless trying to find limitations or implementation constraints, which explain the deviations from optimality might be an interesting way forward for eye movement models, implementing the typical next steps of an optimal observer analysis.

1.5.8 *Systematic tendencies*

Finally, one perspective emphasizes the image independent systematic tendencies in eye movement behaviour (Tatler & Vincent, 2008, 2009). While these accounts do not deny the existence of top-down and bottom-up features guiding eye movements, they add that there are substantial statistical regularities in scanpaths which are not explained by task or stimulus. These tendencies include an initial central fixation bias (Tatler, 2007), an overall coarse-to-fine strategy (Over et al., 2007), a preference for cardinal directions prominent in the image (Foulsham, Kingstone, & Underwood, 2008), a characteristic distribution of saccade lengths (Ramos Gameiro et al., 2017), which depends on the observer (Castelhano & Henderson, 2008), a tendency to return to the previous fixation location, but later than usual eye movements, and a saccadic momentum, i.e. the tendency to continue with eye movements in the same direction after short fixations (Smith & Henderson, 2009). These observations reinforce that fixations should not be analysed independently. Furthermore, some of these tendencies have been implemented into models (Clarke, Stainer, Tatler, & Hunt, 2017; Le Meur & Coutrot, 2016; Le Meur & Liu, 2015), which are highly predictive of eye movement behaviour, suggesting that the current choice of fixation locations depends more on the previous fixation location(s) and duration(s) than on all guidance by image and task combined.

1.5.9 *Difficulty of evaluation*

The large number of approaches results in entirely different evaluation schemes for different types of models, rendering approaches largely incomparable. For models of scanpaths in natural scenes the situation was even more dire, as the evaluation methods here differed even between individual models and laboratories (compare for example Clarke, Stainer, et al., 2017; Clarke & Tatler, 2014; Engbert et al., 2015; Le Meur & Baccino, 2013; Le Meur & Coutrot, 2016; Le Meur & Liu, 2015) and existing methods were not necessarily well justified statistically (Engbert et al., 2015). To solve this problem of inconsistent evaluation, I developed a method to calculate the likelihood for any model which predicts whole scanpaths and assigns a non-zero probability to any measured scanpath. This calculation and the statistical methods enabled by it are described in Chapter 3 and our corresponding publication (Schütt et al., 2017).

1.5.10 *The SceneWalk model*

To illustrate these new evaluation methods I use the SceneWalk model, which was originally developed by our collaborators in Potsdam. This model aims to explain some of the systematic tendencies observed in scene viewing. It takes a desired fixation density as input and predicts scanpaths following this fixation density. This model is also important for this thesis, because the model already provides a method to generate a scanpath based on a fixation density, such that the predicting a fixation density is sufficient to create full scanpaths.

1.6 CONNECTING EARLY VISION AND EYE MOVEMENTS

After developing these new evaluation methods and the image-computable early vision model, the final step of my thesis was to combine these two models into a model which can produce scanpaths for new images.

Because the SceneWalk-model (Engbert et al., 2015) is already able to create reasonable scanpaths for a given fixation density, linking the models to create a combined model requires only the prediction of a fixation density based on the image. Describing the connection as a single, fixation independent saliency map excludes any influence of eye movements on the computation of the saliency map. Removing these dependencies is a simplification as a dependence on the current fixation location is certainly expected, since visual perception declines into the periphery (Strasburger, Rentschler, & Jüttner, 2011).

Furthermore, the connection I propose in Chapter 4 directly predicts a fixation density from the representation created by the early spatial vision model. As I mentioned above, the hypothesis that eye movements are mainly driven by salient low level features has been largely refuted by previous research. Nonetheless I implement such a direct connection, because this allows me to investigate in detail when early visual processing might play a role in eye movement planning and to measure the maximal relative contribution of low-level features. My early vision model implements features which adequately represent early visual processing for the first time and strengthens the conclusions that can be drawn substantially.

To extract more information for the dynamical modelling of eye movements, I evaluate my predictions and other saliency models—including models with higher level features—over time. Ultimately, I'll conclude in accordance with the field that the role of bottom up saliency in static scenes is small and restricted to earliest fixation(s) or not existent at all.

Nonetheless, the predictions based on early visual processing (or another saliency model) can be combined with the SceneWalk model to form a first simple incarnation of the model I originally aimed at, which predicts scanpaths for new, previously unseen images. Based on this proof of concept, more detailed models including higher level features, more complex sequential dependencies and influences of eye movements on the computation of saliency may be developed in the future as I discuss in Chapter 5.

1.7 DATA SOURCE

The evaluation of the models I develop in this thesis requires experimental data and ideally this data should be collected specifically to provide a hard test for the models.

For the early vision model detection and discrimination was abundantly available from the literature and from earlier studies performed in the lab of Felix Wichmann in Tübingen, Berlin and Oxford. Thus, no further data collection was required to design and test the early vision model.

For the eye movement models we even had specifically collected data available to develop my models, due to a collaboration with researchers in Potsdam. Lars Rothkegel pursued an experimental PhD-thesis with Ralf Engbert in Potsdam for which he collected eye movement data, specifically to test hypotheses about the models I developed.

More specifically three important datasets were available, which I used for model evaluation for this thesis. The first is a moderately sized free viewing dataset collected by Hans Trukenbrod (Potsdam), which was available early on during the project. I used this dataset (Engbert et al., 2015) for the development of the evaluation methods I present in Chapter 3. Second, Lars Rothkegel collected a large Corpus dataset of free viewing data, which provides the basis of the more detailed, time resolved analysis of fixation densities I present in Chapter 4. Third, I directly participated in the design of an experiment executed by Lars Rothkegel (Rothkegel et al., 2018), in which observers searched extensively for targets which are commonly used in studies on early visual processing. These search data are the basis of my analyses of task effects in Chapter 4.

1.8 OUTLINE

After the introduction this thesis proceeds with three major content parts before it concludes with a general discussion:

In Chapter 2, I present my image computable early spatial vision model, which implements our knowledge about the first steps of visual processing such that it is applicable to natural images. This is necessary to allow us to make any statements about the stimuli used to measure eye movements. We also published this model in almost identical form in the article (Schütt & Wichmann, 2017).

In Chapter 3, I present my work on the evaluation of eye movement models. Prior to this work dynamical eye movement models were usually fit using ad hoc criteria without much statistical justification. Thus, fitting and evaluation of such models was a tedious and subjective process. Now, in the likelihood based framework I present in this chapter evaluation can be done using a single measure of model fit, which is well justified statistically and allows many more advanced analyses of the model and data. These results were also already published in the article (Schütt et al., 2017).

In Chapter 4, early vision models and eye movement models are united. First, a model is developed to predict the fixation density from the representation produced by the early spatial vision model from Chapter 2. Then this model is evaluated over time on eye movement data obtained with two different tasks: Free viewing and visual search in natural scenes. A corresponding manuscript (Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2018) is currently in preparation.

Finally in Chapter 5, I discuss my overall results, their implications and possibilities for further research.

Numerus hominum, in quibus experimenta instituta sunt.		Differentia minima unciarum vel drachmarum manibus impositarum, in qua diversitas ponderis percipiebatur.			
1.	tactu	32 Unc.	: 17 Unc.	differt	15 Unc.
	tactu et coenaesthesi	32 Unc.	: $30\frac{1}{2}$ Unc.	—	$1\frac{1}{2}$ Unc.
	tactu	32 Drachm.	: 24 Drachm.	—	8 Dr.
	tactu et coenaesthesi	32 Drachm.	: 30 Drachm.	—	2 Dr.
2.	tactu	32 Unc.	: 22 Unc.	—	10 Unc.
	tactu et coenaesthesi	32 Unc.	: $30\frac{1}{2}$ Unc.	—	$1\frac{1}{3}$ Unc.
	tactu	32 Drachm.	: 22 Drachm.	—	10 Dr.
	tactu et coenaesthesi	32 Drachm.	: 30 Drachm.	—	2 Dr.
3.	tactu	32 Unc.	: 20 Unc.	—	12 Unc.
	tactu et coenaesthesi	32 Unc.	: 26 Unc.	—	6 Unc.
	tactu et coenaesthesi	32 Drachm.	: 26 Drachm.	—	6 Dr.
4.	tactu	32 Unc.	: 26 Unc.	—	6 Unc.
	tactu et coenaesthesi	32 Unc.	: 30 Unc.	—	2 Unc.
	tactu et coenaesthesi	32 Drachm.	: 29 Drachm.	—	3 Dr.

Weber (1834)

In this chapter I present my image-computable implementation of the standard early spatial vision model. This model will be used in Chapter 4 to analyse the influence of early visual processing on eye movements. Both content and form of this chapter equal our recent article:

Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, 17(12), 12:1-35.

2.1 INTRODUCTION

The initial encoding of visual information by the human visual system has been studied extensively since the seminal studies of the late 1960s and early 1970s (e.g. Blakemore & Campbell, 1969; Campbell & Robson, 1968; Carter & Henning, 1971; Graham & Nachmias, 1971; Nachmias & Sansbury, 1974). Their insights have shaped how we now think about the first computations of the visual system: spatial frequency and orientation specific channels followed by a static nonlinearity. This conceptual model is both broadly consistent with physiology up to primary visual cortex, as well as with normative theories on how the available information should be processed.

As a conceptual framework, the standard model of spatial visual processing is useful and successful. Computational models of it, however, are usually only implemented to work with an abstract representation of visual stimuli, not with "real" images. Typically, the models start with activity in the frequency channels, calculated—or taken—from the parameters of the simple one-dimensional stimuli (e.g. Foley, 1994; Goris et al., 2013; Itti et al., 2000; Legge, Kersten, & Burgess, 1987). This simple implementation of early spatial vision models is highly efficient because first, it bypasses the computational intensive multiscale image decomposition and second, it requires few computational units because it is only one-dimensional (1D)—the models are only applicable to (simple) one-dimensional stimuli. Historically, it was the lack of computational power which precluded image-computable models.

Implementing a model to be image-computable, i.e. to work on any image as input, helps to generalise its application to a wide range of tasks and datasets—only image-computable models allow quantitative predictions for any input image (c.f. the discussion of the importance of image-computability by D. L. Yamins & DiCarlo, 2016, in the context of convolutional deep neural networks (DNNs) as models of object recognition). Furthermore, image-computable models may reveal—and make it easier to explore—potentially counter-intuitive effects of nonlinearities in one's model. Another benefit is that image-computable models of early spatial vision may be useful beyond spatial vision, because they can be used as psychophysically plausible preprocessors in investigations of higher level processing and for more natural tasks. Finally, image-computable models allow the investigation of statistics of the model *output*, comparing it to normative theories from, e.g. the efficient coding hypothesis (Attneave, 1954; H. B. Barlow, 1969; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001; Simoncelli & Olshausen, 2001).

But even for spatial vision, an image-computable model may aid further development: an image based implementation necessarily requires that the model is implemented in full 2D, including orientations and the spatial sizes of filters and normalization pools; they necessitate to think about spatial vision jointly in the space as well as the spatial-frequency domain. This aspect is likely important for the understanding of visual processing (Daugman, 1980), but is typically not implemented in the abstract, 1D models (Goris et al., 2013).

In this chapter I present a psychophysical, image-computable model for early spatial visual processing; I aim to explain human performance in behavioural tasks and thus evaluate my model only on behavioural data from human observers.

2.1.1 *History and classical experiments in spatial vision*

Psychophysics has a long tradition of quantifying behaviour summarizing it using equations—often called "laws" to mimic physics (Fechner, 1860; Stevens, 1957; Weber, 1834). We have a good quantitative understanding of sensitivity to luminance differences, the dependence of luminance discrimination on wavelength, and the size of test patches (reviewed by Hood, 1998; Hood & Finkelstein, 1986). These early results allow us to convert physical light patterns first into luminance patterns and subsequently into contrast images. The contrast images largely determine detection and discrimination performance (once the display is sufficiently bright).

Arguably, the advent of modern spatial vision came with the discovery of spatial frequency and orientation tuned "channels" (Campbell & Kulikowski, 1966; Campbell & Robson, 1968). Later, the existence of these channels was confirmed by numerous studies, including signal mixture and adaptation experiments (e.g. Blakemore & Campbell, 1969; Graham & Nachmias, 1971). The postulate of independent spatial frequency and orientation channels allows to predict detection thresholds for any signal pattern from the knowledge of the Fourier spectrum of the stimulus and the sensitivity to single sinusoidal gratings of different frequencies, i.e. the contrast sensitivity function.

Because of its pivotal role in the early linear channel model, the contrast sensitivity function was measured under many different conditions, including peripheral presentation (Baldwin, Meese, & Baker, 2012; Rovamo & Virsu, 1979; Virsu & Rovamo, 1979), different luminances (Hahn & Geisler, 1995; Kortum & Geisler, 1995; Rovamo, Luntinen, & Näsänen, 1993; Rovamo, Mustonen, & Näsänen, 1994), different temporal conditions (Kelly, 1979; Watson, 1986; Watson & Nachmias, 1977) and different spatial envelopes (Robson & Graham, 1981; Rovamo et al., 1994).

Another line of research investigated how the (putative) spatial frequency channel responses are further processed and combined to produce visual behaviour. This line of research started with contrast discrimination experiments, measuring the contrast increment needed in addition to a pedestal contrast to produce a detectable difference (Foley & Legge, 1981; Nachmias & Sansbury, 1974). Typically the so-called "dipper function" is found: low pedestal contrasts facilitate detection, i.e. discrimination can be better than detection, while discrimination requires progressively larger contrast increments for growing pedestal contrast (as to be expected from Weber's law). To explain the shape of the dipper function, Legge and Foley (1980) proposed a Naka-Rushton nonlinearity (Naka & Rushton, 1966) on the spatial frequency channel outputs. Later Foley (1994) revised this model to replace the single-channel nonlinearity with a normalization by the other channel responses to explain oblique masking data, i.e. experiments in which the mask grating had a different orientation than the signal to be detected. This across-channel-normalization is in spirit very close to the typical divisive contrast-gain control introduced to explain the behaviour of simple cells in V1 (Cavanaugh et al., 2002a; Geisler & Albrecht, 1995; Heeger, 1992).

Finally, the last processing step of (most) models in vision is one of decoding: deriving the open behavioural response from the activity in the model. In older spatial vision models simple task-independent Minkowski norms were used (the popular "max-rule" or "winner-takes-all-rule", i.e. the decision is based on the maximally active unit or channel

only, corresponds to a Minkowski norm with large—in the limit infinite—exponent). Decoding as an important part of spatial vision models was first discussed by Pelli (1985) in the context of uncertainty. In more modern models, channels are explicitly modelled to respond noisily such that the decoding can be understood in its original statistical meaning of deriving the response from the noisy channel responses. Frequently, this decoding is assumed to be optimal (e.g. Goris et al., 2013; May & Solomon, 2015a, 2015b).

Much of the history of the field, its psychophysical experiments and the purely abstract 1D spatial vision models are summarised and discussed in the comprehensive book of Graham (1989).

There have been earlier attempts to make image-computable models of spatial visual processing, for example by by Teo and Heeger (1994) and by Watson and Solomon (1997). However, these earlier models were limited by the available computational power at their time, which required them to tailor their models to the processed stimuli or to limit the possible computations, for example to entirely local normalization. Recently some more models were implemented to work on images (e.g. Alam, Patil, Hagan, & Chandler, 2015; Bradley et al., 2014). These models usually do not cover the whole complexity, but simplify the normalization steps to reach a computationally more efficient model (Bradley et al., 2014) or are based on entirely different approaches like neural networks trained to predict the detectability of specific distortions (Alam et al., 2015).

One mayor incentive to develop image computable models of early visual processing are the applications in image processing. The classical aim here is image quality assessment, i.e. to produce a metric which measures how bad a particular distortion of an arbitrary image is as perceived by humans. Consequentially, the classical models were immediately proposed as such image quality metrics (Teo & Heeger, 1994; Watson, Borthwick, & Taylor, 1997). Such an image quality metric can then be used to optimize various image processing algorithms like compression or tone mapping. This cascade towards application has recently been demonstrated for a different biologically inspired model, the normalized Laplacian pyramid (Laparra et al., 2016; Laparra, Berardino, Ballé, & Simoncelli, 2017). My model seems to be a good start for a similar path towards application as it makes valid predictions what distortions are visible to humans and also the optimization of supra-threshold distortions yields reasonable predictions as we shall see below.

2.1.2 *Outline*

In the following, I first describe how I implemented the spatial vision model to operate on images. I then show that my model reproduces classical psychophysical spatial vision findings, namely those which gave rise to the now accepted model structure in terms of linear filters and divisive normalization. Thereafter, I evaluate the model on a dataset measuring masking by natural images. Then I show that the model produces a sparse representation, as predicted—and desired—from normative considerations. As a final step, I create optimized stimuli to maximize or minimize differentiability according to the model.

2.2 MODEL DESCRIPTION

Like most image processing spatial vision models, my model contains 4 major parts: Images are first *preprocessed*. Then they are *decomposed* into spatial frequency and orientation specific channels and pass an accelerating *nonlinearity and normalization*. Finally,

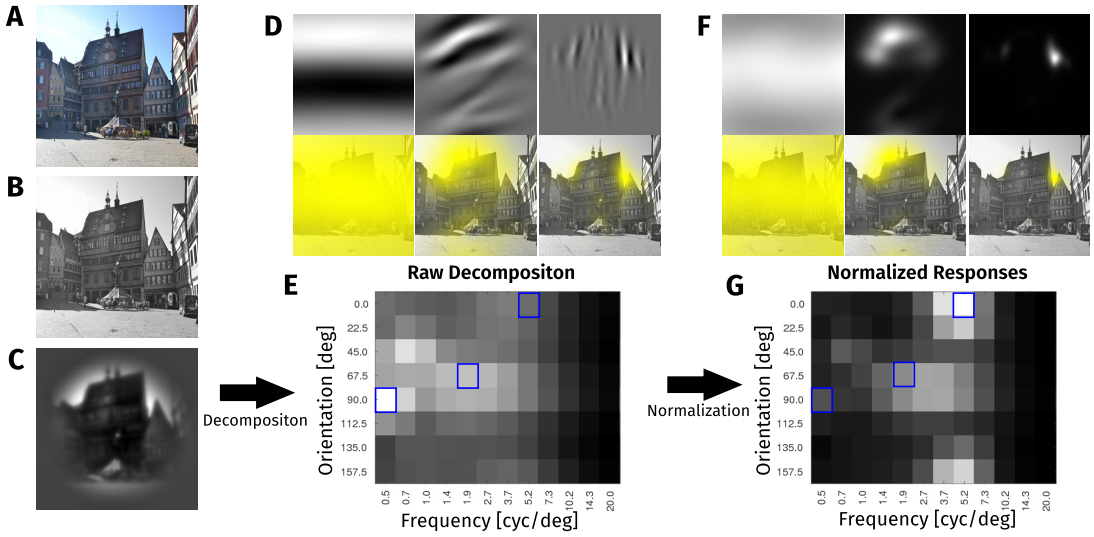

Figure 1: Overview over the model processing. As an example, a photograph of the town hall of Tübingen (**A**) is passed through the model. **B** shows the image after conversion to luminance and **C** shows it after incorporation of eye optics, a hand-tuned contrast sensitivity function and cut-out of the fovea. The image is then decomposed into spatial frequency and orientation channels. The output of these channels for the example image is displayed in **D** and **E**. **D** shows the real part of the output and the absolute value of the output overlayed on the image for three example channels marked in **E**. **E** shows the mean absolute value of each channel. Finally, each channel's activity is passed through an accelerating nonlinearity and is normalized by a surrounding normalization pool. The result of this is displayed in **F** and **G**. **F** shows the activity of the same three channels as **D** after normalization, first isolated and then overlayed over the original image. **G** shows each channels' mean activity over the image after normalization.

for *decoding*, I assume additive noise and optimal decoding to predict how well images can be differentiated.

2.2.1 *Preprocessing*

In most psychophysical experiments, stimuli are directly defined in contrast units because the pattern and the contrast together explain most variance, once the stimuli are bright enough. Thus, these stimuli could be passed into my model as they are defined, without any preprocessing.

Nonetheless, I implement the conversion from physical light patterns into the contrast coded input to the main processing explicitly for two reasons: First, I aim for a model, which can process arbitrary images displayed on a screen and images are usually not given in contrast units (as the example image in Fig. 1 A). Second, optical effects and retinal processing could be modelled in more detail than I do here. Thus the simple preprocessing steps mark, where in the model more complex precortical processes fit in and which properties of them are modelled.

First, all images are converted to luminance values at each pixel. The stimuli used in the classical experiments were already given in luminance values. For modelling the natural image masking database by Alam et al. (2014) as described below, I use the pixel value to luminance conversion function as provided with the data. For all other natural images, I used measured spectra from a monitor in the lab in Potsdam (Mitsubishi Diamond Pro 2070) and the V_λ curves as given by Sharpe, Stockman, Jagla, and Jägle (2005) to convert the pixel values to luminance. This monitor was used for the eye movement experiments I use for the evaluation of the models responses below. For display they were converted back to RGB values by calculating the nearest value with equal strength in all 3 channels (See Fig. 1 B for an example).

Next, I apply optical distortions according to the mean modulation transfer function of a well corrected human eye. To do this, I use a formula by Watson (2013), which was based on optical aberration measurements by Thibos, Hong, Bradley, and Cheng (2002) on 200 eyes of 100 healthy, well-corrected subjects. I fixed the pupil diameter required for these formulas at 4 mm for my simulations. The pupil diameter could be measured, experimentally controlled, or estimated from the luminance over the visual field (Watson & Yellott, 2012). However, in none of the experiments fitted here pupil diameter or luminance were varied explicitly and conditions were reasonably similar in all experiments, such that I opted for this slight simplification.

Conversion of stimuli to contrast is then performed by dividing by their mean and subtracting 1. Then the stimulus was cropped to an area of $2° \times 2°$ of visual angle around the assumed fixation location (for most classical stimuli the center of the stimuli where they reach maximal nominal contrast). If the stimulus was smaller than $2° \times 2°$, I filled the rest of the area with zeros. Finally, I resized the image to 256×256 pixels using MATLAB's "imresize" function, which performs a bicubic interpolation.

I then implement the higher neuronal sensitivity for medium to high spatial frequencies as an additional linear filter similar to the "high pass filtering of neural origin" of Rovamo et al. (1993), which depends on presentation time. As in earlier approaches, I estimated the neuronal influence on contrast sensitivity simply as the necessary filter to match contrast sensitivity. To implement this filter with as few assumptions as possible, I fitted its modulation transfer function (MTF) by hand as a third order spline.

To complete preprocssing, I smoothly cut out the image patch corresponding to the fovea, as I want to restrict myself to foveal processing here and to avoid any border effects

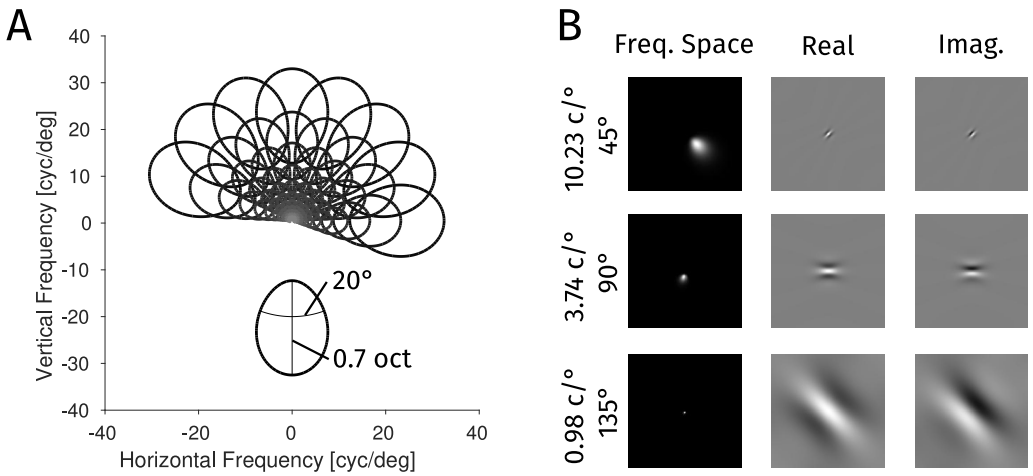

Figure 2: Illustration of the filters used for the decomposition. **A**: half response curves in frequency space for all filters. Lighter grey for higher frequency channels. Additionally, one filter is displayed separately to show the half bandwidths at half height of each channel. The distribution of channels may appear tilted in the figure, because I included filters in the cardinal directions; however, by mirror symmetry the filters cover or tile the space equally. **B**: Three example channels of different frequency and orientations relative to horizontal. For each channel, a heat map of the weights in frequency space and the real and imaginary part of the filter weights in space are given. The similarity of the filters to receptive fields of V1 neurons is not incidental.

in later processing. For this purpose, I use a $2° \times 2°$ raised cosine window. This window is above half height over the central disc of $1°$ diameter, roughly fitting the size of the foveola with maximal resolution and sensitivity.

The final preprocessing result for an example image is displayed in Figure 1 C.

2.2.2 *Decomposition*

Next I aimed to implement the well established orientation and spatial frequency selective channels (Campbell & Robson, 1968). These were implemented as a dense filter-bank with each individual filter fitting psychophysical and neuronal measurements of channel specificity, as illustrated in Figure 2.

Many functional forms exist that can represent the filter shape of the psychophysical channels closely enough. Here I chose to use a log-Gabor as the basic filter shape, which corresponds to a Gaussian shape in log-frequency and in orientation. A log-Gabor is directly and completely defined by its preferred spatial frequency and orientation and its bandwidth in each dimension, which are all properties estimated from psychophysical and physiological data routinely. Additionally, Gabor-filters are maximally localized jointly in space and frequency, have a monotonically and smoothly decreasing response for frequencies and orientations moving away from the preferred parameters and no response to uniform fields. These are all desirable properties for a sub-band decomposition, which gives my filter choice some normative justification. Ultimately however, any functional form that closely represents the specificities of the psychophysical channels (and thus, V1 neurons) will yield indistinguishable responses in the channels and thus results indistinguishable from my choice.

Additionally to spatial frequency and orientation specificity, linear filters are also tuned to the phase of the stimulus as simple cells in primary visual cortex are (Daugman, 1980). However, psychophysical performance seems not to depend on absolute phase. The most parsimonious model to achieve such phase independent behaviour is to use a quadrature pair, i.e. filters which differ only in their phase preference and exactly by 90°. Such a quadrature pair is usually written as a single complex filter with one filter defining the real and one defining the imaginary part of the filter optimizing the implementation further. From a quadrature pair, the response of a filter preferring any phase can be computed as a linear combination of these two filters. Especially, the absolute value of the complex response can be computed, which represents the response of an optimally phase-tuned channel at each position. For my channels I implemented this scheme and pass only the absolute value of each channels' response on to further processing, as illustrated in Figure 3. As I demonstrate in Figure 3 B, this treatment of phase indeed leads to a phase independent response.

Quadrature pairs could be implemented neuronally using four phase preference types of neurons for positive and negative responses of the two filters in the pair as discussed by Watson and Solomon (1997). Indeed, neurons in macaque primary visual cortex cluster around even and odd symmetric phases (Ringach, 2002). However, there are neurons at all preferred phases and strongly orientation tuned neurons tend to prefer odd phase while less tuned neurons tend to prefer even phase. Both of these observations are incompatible with a direct implementation of quadrature pairs in neurons. Consequently quadrature pairs must be seen as a simplification.

I set the bandwidth of the channels based on the literature, as I do not include data here that could constrain the spatial frequency selectivity of the channels. For spatial frequency, I chose a standard deviation σ_F of 0.5945 octaves corresponding to 0.7 octaves half bandwidth at half height, roughly matching the adaptation data of Blakemore and Campbell (1969) and the neural data of Ringach, Shapley, and Hawken (2002). For orientation, I chose a standard deviation σ_θ of 0.2965, corresponding to 20° half bandwidth at half height based on early psychophysical measurements (Campbell & Kulikowski, 1966; Phillips & Wilson, 1984). These measurements used oblique masking data to estimate the bandwidth of the channels, similar to some data I present below. Consequently, any substantial deviation of the estimates would be noticeable when comparing my predictions to these data. Additionally, these estimates are in good agreement with physiological measurements (Campbell, Cleland, Cooper, & Enroth-Cugell, 1968), as already noted in the original papers and do fit more modern measurements like Ringach et al. (2002). Nonetheless, my filter collection only roughly approximates the neural population, because there is substantial variability in the specificity of cortical neurons (Goris, Simoncelli, & Movshon, 2015; Ringach et al., 2002) and I ignore known dependencies between preferred spatial frequency and the bandwidths (Phillips & Wilson, 1984), an issue on which I comment in more detail in the discussion of this chapter.

Finally, it needs to specified how many channels at which spatial frequencies and orientations to use. Normative theory from signal processing tells us that two different orientations and octave spaced spatial frequency channels suffice to represent the whole information present in a image as it is done for wavelet decompositions (Strang & Nguyen, 1996). Commonly, pyramid schemes are applied to achieve such a decomposition with as few filter responses and as little computation as possible (Simoncelli, Freeman, Adelson, & Heeger, 1992; Watson, 1987). Specific types of filters allow these pyramids to achieve additional advantageous properties like steerability or shiftability (W. T. Freeman & Adelson, 1991; Simoncelli et al., 1992).

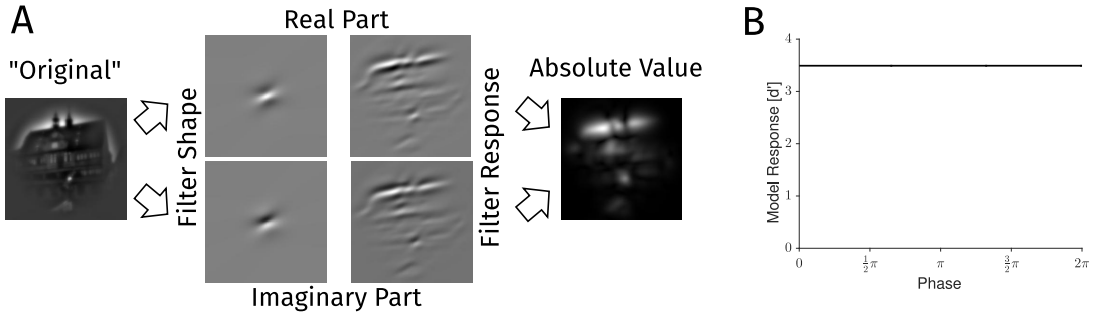

Figure 3: Illustration of the phase handling in the model. **A**: The preprocessed example image called "Original" passes through the processing of a single channel. The complex filtering corresponds to filtering with two filter shapes, an even phase filter for the real part and an odd phase filter for the imaginary part, which are illustrated in the second column. In the third column the responses of the filters to the image are shown, which are then combined to the absolute value at each position illustrated in the last panel. **B**: The model response (plotted as the signal to noise ratio d' for detection of the stimulus) to a $3 \times 3°$ Hanning windowed horizontal grating of 10 $\frac{cyc}{deg}$, changing the phase of the grating. The response of the model is phase independent up to numerical precision.

To achieve this, however, one needs to choose specific filter shapes which need to be broad in frequency and orientation. Using narrower filters, more different filters are required to cover all orientations and there are only discrete choices which fix both bandwidth and number of channels in each scheme. Even worse, for spatial frequency the whole pyramid scheme breaks down once one wants channels that are not octave-spaced because downsampling by other factors than two is much less efficient. Thus, these pyramid schemes do not allow us to fit the channel bandwidths and the density of channels independently and limit us to octave spaced channels.

One could glance over this and approximate the filters with the best fitting pyramid as Watson and Solomon (1997) did for example, if there was nonlinear processing after the decomposition. A stimulus that matches a filter in the model leads to a single large response in that channel, while a stimulus between channels leads to several smaller responses. Then the accelerating nonlinearity amplifies the larger response more than the several smaller responses leading to a stronger model response to stimuli that match a channel than to stimuli that fall between channels.

In my model this leads to an oscillating response with peaks at the orientations and spatial frequencies of the channels (see Figure 4). Note that such oscillatory behaviour must occur for any model, that employs nonlinearities after the decomposition in channels for specific frequencies and orientations. A nonlinearity imposes different weights on the channels depending on signal strength. However, the activities of any set of linear channels keep the same relative strength when the absolute signal strength changes. Thus, no linear channel shape can fully compensate for the nonlinearity unless the nonlinearity is computing energy, i.e. squaring and summing over channels.

In contrast, one observes neither oscillating performance nor any clustering in preferred spatial frequency or orientation in either psychophysics or neurophysiology. Neurons seem to cover every frequency and orientation in the range they cover and human performance on psychophysical tasks seems to change smoothly with scale and orientation.

To mimic the dense neural covering of spatial frequencies and orientations, I chose to simply increase the number of frequency and orientation channels until the oscillations of performance were sufficiently small (see Figure 4). This method allows us to keep the

implementation as a convolution, which is still necessary to reach an acceptable computation time. An implementation that includes a realistic sampling of the channels would go far beyond my horizon here as this seems not to be constrained psychophysically and such decompositions with variable channels were not studied in detail so far.

Following these considerations, I used a complex-valued log-Gabor filterbank with 8×12 filters for orientation and spatial frequency for my decomposition. The 8 preferred orientations were equally spaced over 180° covering half the frequency space. The 12 preferred spatial frequencies were placed logarithmically on the spatial frequency axis from $0.5\frac{cyc}{deg}$ to $20\frac{cyc}{deg}$, which roughly covers the range of frequencies visible to human observers. The kind and range of filters I used are illustrated in Figure 2.

Each of the filters was precomputed in frequency space. I then calculated the filter response by multiplying the Fourier transform of the preprocessed image with the frequency space representations, which yields a complex-valued image each. This complex-valued image contains responses of an even symmetric filter as its real part and the responses of an odd symmetric filter as its imaginary part. As discussed above, I pass the absolute value of this response on to further processing, dropping phase entirely.

The results of the whole decomposition stage are illustrated for the example natural image in Figure 1 D and E. In D I show the results before and after removing phase information for 3 example channels. In E you find an overview over all channels in which I display only the average absolute response of each channel.

2.2.3 *Normalization & nonlinearity*

Masking and contrast discrimination experiments show clearly nonlinear relationships between thresholds and mask contrast (Legge & Foley, 1980). To model these psychphysicyal results and the corresponding interactions observed in primary visual cortex neurons (Cavanaugh et al., 2002a; Heeger, 1992), the channel activities are passed through a divisive normalization (Carandini & Heeger, 2012; Foley, 1994; Heeger, 1992; Watson & Solomon, 1997). In my model, I restrict the normalization to a pool localized in space, spatial frequency and orientation. The localization in space and frequency is not controversial, while it is sometimes claimed that the normalization pool is not orientation selective, on which I comment in the discussion of this chapter.

In older models, this step was modelled as a Naka-Rushton nonlinearity (Foley & Legge, 1981; Legge & Foley, 1980; Naka & Rushton, 1966), which is equivalent to this normalization with an extremely narrow pool that contains only the channel itself as an input.

In my model the formula for divisive normalization of original channel activities $A = (a_i)_{i\in\mathcal{I}}$ to compute normalized final responses $R = (r_i)_{i\in\mathcal{I}}$ is:

$$r_i = \frac{a_i^{p+q}}{C^p + b_i} \tag{1}$$

Using an index set $\mathcal{I}$, which indexes all different channels and all positions, a constant C, exponents p and q and $B = \{b_i\}_{i\in\mathcal{I}}$, an array of normalization coefficients, which are computed from the element wise powers $A^p := (a_i^p)_{i\in\mathcal{I}}$:

$$B = A^p * G \Leftrightarrow b_i = \sum_{j\in\mathcal{I}} G(x_i - x_j)a_j^p, \tag{2}$$

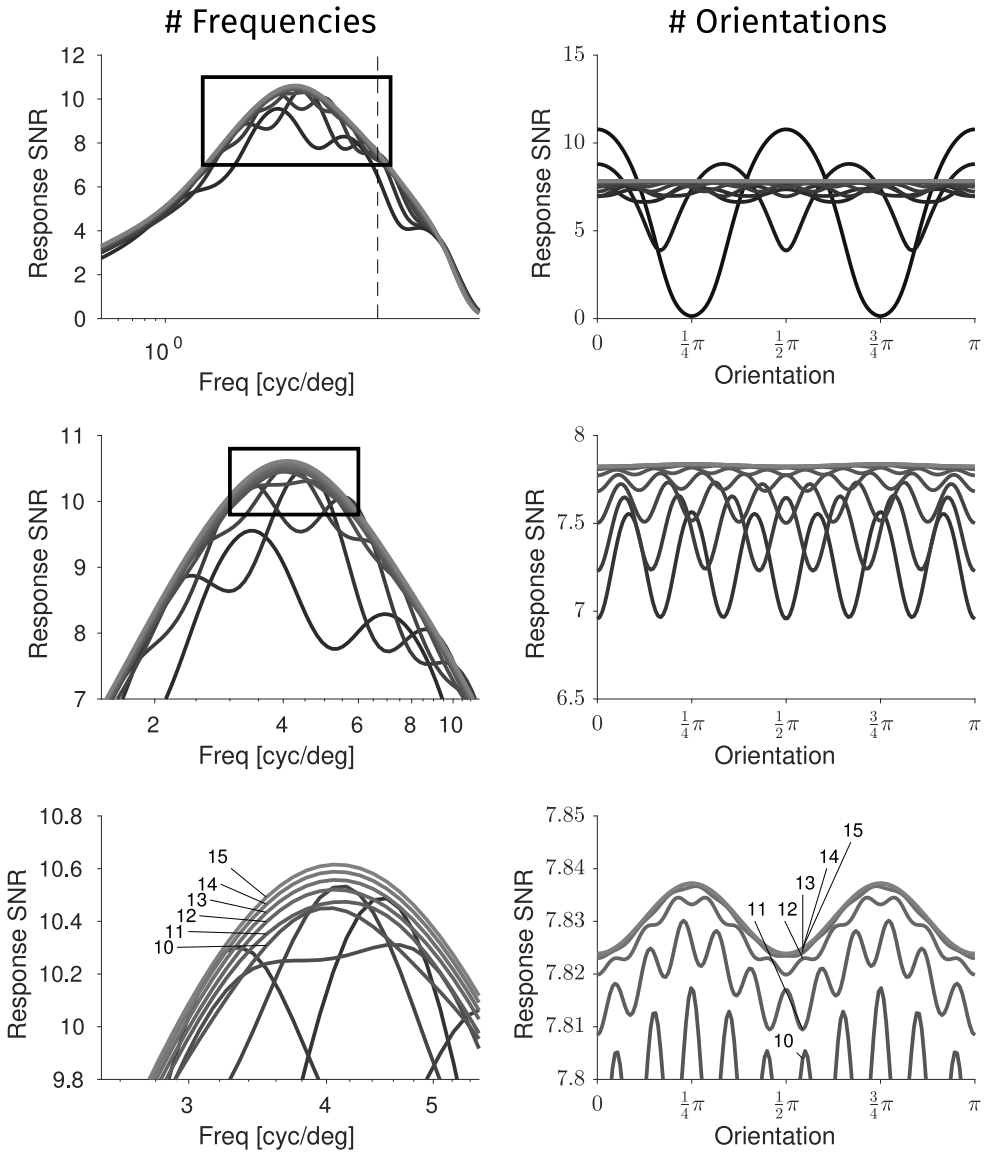

Figure 4: Illustration of the effects of using fewer channels on model performance. Left column: Estimated signal to noise ratio for a $3 \times 3°$ Hanning-windowed horizontal grating with 10% contrast against the frequency of the grating. Lighter grey levels correspond to more channels. Each row shows the marked area in the row above, representing different zoom levels. Right column: As the left column, but fixing the frequency of the grating to $10\frac{cyc}{deg}$ (marked in the left column by a dashed line) and varying the orientation of the grating instead. The results shown here were obtained for 256x256 images, but the effects are largely independent of image size

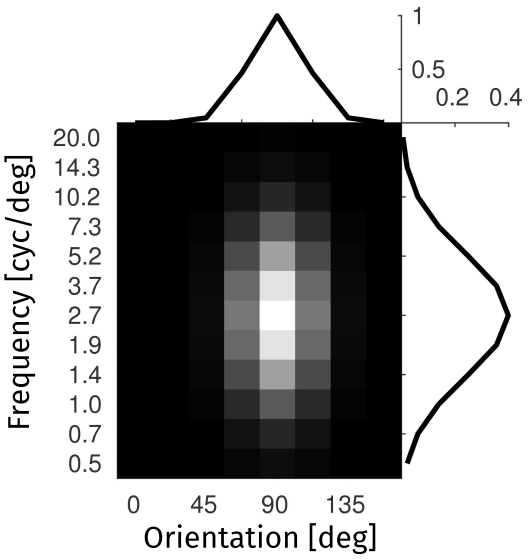

Figure 5: Illustration of the normalization pool G over spatial frequency and orientation. Both are shown for the central pixel in the $2.7\frac{cyc}{deg}$, 90° orientation channel

by convolution with G, a 4D Gaussian normalization pool with standard deviations $\omega_x = \omega_y$ in space, ω_F in spatial frequency and ω_θ in orientation.[1]

The weights for the pool in spatial frequency and orientation are displayed in Figure 5. For frequency I set the size of the normalization pool to a rough estimate of $\omega_F = 1$ octave standard deviation. For orientation I fit the pool bandwidth ω_θ based on oblique masking data (displayed in Figre 11), as explained in more detail below.

For the spatial extent I first implemented the model using a Gaussian profile. However, I lack the data to constrain the size of the normalization pool in space. Instead of arbitrarily setting a pool size, I tested the extreme cases of such a model here. Specifically, I set the normalization pool to be either only the exact pixel to be normalized or all responses over the image weighted equally. These cases correspond to an infinitely small and a infinitely large pool respectively. For the classical grating based data, I find that the normalization over the whole image leads to a better result and more consistent parameter estimates, while the natural image data is better explained by the perfectly local normalization.

Nonetheless, I neither believe that the normalization pool is perfectly local nor that is fills the whole space. Both psychophysical (Snowden & Hammett, 1998) and neural data (Cavanaugh et al., 2002a) suggest that the normalization pool has some extent beyond the classical receptive field (roughly 2.5-3 times the radius from the neuronal data). Also the model allows arbitrary intermediate sizes for the normalization pool and sporadic fits I made with intermediate pool sizes yielded good fits to the classical data as well. Consequently, I do not argue against the normalization pool having a non-zero spatial extent.

The additional exponent q is required, because a single saturating function per channel cannot explain the discrimination thresholds at high contrasts, which grow much less than predicted from a saturating response function (Goris et al., 2013). This approach was used earlier by Foley (1994) and Watson and Solomon (1997) in their models.

1 Technically, I should use a von Mises distribution for orientation, which wraps the tails of the normal distribution around as orientation is a circular dimension. However, as the normalization pool I find is narrow, the difference between a Gaussian and the von Mises distribution is negligible.

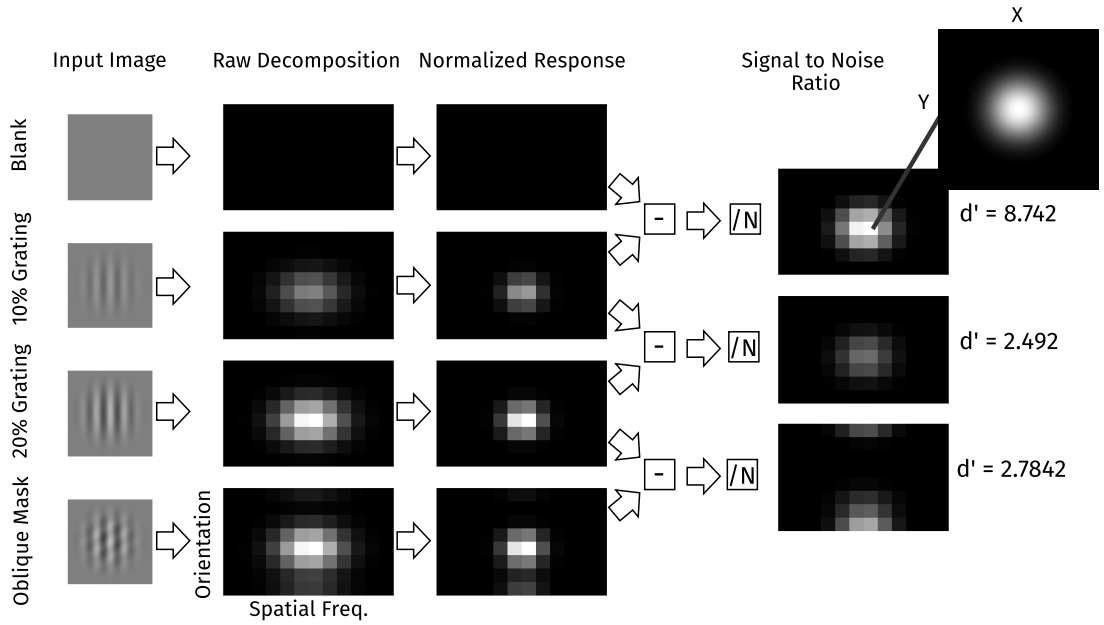

Figure 6: Illustration of the read-out mechanism of the model. For four different typical spatial vision stimuli, I show the channel mean of the raw decomposition results and of the final normalized responses. To predict how well two images can be differentiated in psychophysical experiments the responses for each pixel in each channel are subtracted from each other and divided by the noise standard deviation. This results in a signal to noise ratio for each position in each channel indicating how well this pixels' activity differentiates the two images. The mean of these signal to noise ratios over each channel are shown in the last column. For one channel I also show the spatial distribution of the differentiability and for each comparison I report the overall discriminability d'. The three pairs of stimuli correspond to contrast detection, contrast discrimination and oblique masking experiments, respectively.

The neural mechanism allowing high contrast discrimination with saturating neurons seems to be neurons with higher C, which start to respond only at higher contrasts. Following Watson and Solomon (1997), I interpret the function in (1) as the sum of responses of neurons responsible for different contrast ranges. For such a sum, the formula with $q > 0$ is practically equivalent as Watson and Solomon (1997) discuss in detail (see their Figure 16 and Discussion point 4.E). As I am not aware of any psychophysical data requiring a separation into contrast channels, I do not include this complication here.

The results after the nonlinearity and normalization are displayed for a natural image in Figure 1 F and G. As for the raw decomposition in D and E, the spatially resolved responses for three example channels are displayed in F and the average response for all channels in G.

2.2.4 *Noise and decoding*

Finally, I need a method to quantify how well stimuli can be discriminated based on their model representations. Here I model noise on the channel outputs and then assume that the rest of the brain optimally decodes from the noisy channel outputs. This allows us to predict whole psychometric functions, i.e. how the proportion of correct responses grows with growing differences. Additionally, it provides a more plausible mechanical interpretation than just computing the difference and pooling with some Minkowski norm

as done by earlier models. The computations for the decoding are illustrated for some typical spatial vision stimuli and tasks in Figure 6.

For my model, I assume independent Gaussian noise for each individual pixel in each channel whose variance scales linearly with the activity in the channel. This model allows us to scale smoothly between pure constant noise and noise that scales completely with the response. Obviously, the independent Gaussian is a specific choice. However, the decision variable will be roughly Gaussian distributed whatever the original distribution was, as the decoding combines many responses for any decision. I also include no noise correlations here, as it would impose a high computational hurdle and is most probably not constrained by the psychophysical data. I discuss my choice of noise in some more detail in the discussion.

Using this noise model, I can compute a signal to noise ratio for each pixel's ability to discriminate a pair of images. Finally I combine the information using optimal linear decoding, which boils down to a weighting by the signal to noise ratio, as the pixels are modelled as independent.

First, we calculate the variance of the Gaussian noise n_i for any response r_i of the model:

$$n_i = N_c + N_f r_i \tag{3}$$

using two parameters, the variance of a constant noise source N_c and the factor for the linear noise N_f. When fitting to data, I found that q and N_F can compensate each other, such that I set $N_F = 0$ regressing to constant noise below (see Appendix A.1 for details on this).

For the i-th pixel we can then calculate the signal to noise ratio for differentiating two images (1) and (2) from the model responses $r_i^{(1)}$ and $r_i^{(2)}$ at this pixel:

$$s_i = \frac{(r_i^{(1)} - r_i^{(2)})}{\sqrt{n_i^{(1)} + n_i^{(2)}}} \tag{4}$$

Using this signal to noise ratio we can calculate the mean value d_i and variance η_i for each pixel weighted by its signal to noise ratio for discriminating this specific pair of images:

$$d_i = s_i(r_i^{(1)} - r_i^{(2)}) \qquad \eta_i = s_i^2(n_i^{(1)} + n_i^{(2)}) \tag{5}$$

From that we arrive at the summed signal d and its variance η and can calculate the percent correct p_c' for a 2AFC task using the standard cumulative normal distribution Φ:

$$p_c' = \Phi\left(\frac{d}{\sqrt{\eta}}\right) = \Phi\left(\frac{\sum_{i\in\mathcal{I}} d_i}{\sqrt{\sum_{i\in\mathcal{I}} \eta_i}}\right) = \Phi\left(\frac{1}{\sqrt{\sum_{i\in\mathcal{I}} \eta_i}} \sum_{i\in\mathcal{I}} s_i(r_i^{(1)} - r_i^{(2)})\right) \tag{6}$$

Note that this system applies only for exactly two images to be compared. If one wanted to decode information about groups of stimuli the optimal decoder is almost always more complex.

For the natural images I once chose a simpler decoding principle. The simpler decoder weights all pixels and channels equally, i.e. (5) is replaced by $d_i = |r_i^{(1)} - r_i^{(2)}|$ and $\eta_i = n_i^{(1)} + n_i^{(2)}$. This essentially assumes that the decoder weights all channels in the correct direction, but has no information on how well each channel discriminates.

Finally, to handle rare lapses of subjects, I simulate a lapse rate of 1% by rescaling p_c' into the final p_c

$$p_c = \lambda + (1 - 2\lambda)p_c' \tag{7}$$

with $\lambda = 0.005$. Taking these lapses into account is necessary as a predicted p_c of 1 renders failures impossible. Thus, without a modelled lapse rate, lapses at high stimulus levels can strongly influence parameter estimates (Wichmann & Hill, 2001).

2.2.5 *Calculating thresholds*

My model calculates percent correct for differentiating two images. Thus, I require a method to calculate thresholds. I chose to calculate thresholds by a bisection method.

The method starts by testing whether the model predicts observers to be correct at maximal displayable contrast (one minus the mask contrast) with a probability higher than a threshold (typically 75%). If this is the case, the bisection method is started with 0 contrast and the maximal displayable contrast defining the first interval.

In each step of the bisection method, one calculates the predicted percent correct for the center of interval calculated so far and takes this point as the new top or bottom end of the interval depending on whether the predicted percent correct is larger or smaller than the threshold percent correct.

I repeat bisection method steps until the width of the interval divided by the lower end is less than 5% and use the center of the last interval as the threshold estimate.

2.2.6 *Parameter fits*

I fixed the model up to the decomposition into different spatial frequency channels without free parameters. After this however, there are some parameters that need to be fit to data. Namely the two exponents p and q, the constant of the normalization C, the bandwidth of the normalization pool ω_θ and the noise strengths N_C and N_F.

To fit parameters, I calculated a single maximum likelihood fit to the data obtained from all observers. This adequately weights the different datasets available for estimating parameters and uses all of the data well.

In short, I started with a grid search over the unset parameters. As a conclusion from this grid search, I restricted myself to a purely constant noise source setting N_F to zero, because I found that changing q can fully compensate for different N_F, such that the model can explain the data equally well, largely independent of N_F. Additionally, I fixed the bandwidth of the normalization pool ω_θ based on the oblique masking data starting an optimization of this parameter from the grid search result.

Using the fixed normalization bandwidth and the purely constant noise source, I then fitted the other parameters to the contrast discrimination data for each presentation time and once additionally for the oblique masking data. For this fitting step I used a quasi-Newton optimization.

Additionally, I decided to fit the parameters again for the ModelFest dataset. As this dataset contains only threshold data I had to convert these thresholds into contrast, percent correct pairs for fitting. If I used only a data point at threshold this favoured shallow psychometric functions that predict threshold percent correct for any pairs of stimuli. To avoid this I added a data point at 1.5 times threshold contrast with 199 of 200 trials correct and a data point at a third of the threshold with 100 of 200 correct trials representing chance performance. As threshold detection data usually does not constrain the normalization exponent q, I fixed it to the value from the longest presentation time of $1497ms$. Fits with this parameter free yielded similar prediction quality.

I give a more detailed description of the fitting method in Appendix A.1.

2.2.7 *Data for model evaluation*

The data for contrast detection, contrast discrimination and oblique and plaid masking were collected during the doctoral studies of Wichmann (1999). Some of the data are published in Bird, Henning, and Wichmann (2002). In these reports all technical details can be found and I report only an overview here.

The classical psychophysical data were collected as temporal two alternative forced choice (2-AFC) experiments, i.e. two stimuli were presented in succession and the observers' task was to report which time interval contained the signal. Presentation time was marked with tones and there was immediate auditory feedback indicating which was the correct interval. In total, 7 observers participated, who were all experienced psychophysical observers, were aware of the purpose of the experiments and had normal or corrected to normal visual acuity. Stimuli were presented on a calibrated, digitally linearised CRT screen with a mean luminance of $88.5\frac{cd}{m^2}$ with a refresh rate of 152.3 Hz. To guarantee independence of signal and mask in the stimuli, they were presented in different refreshes combining 3 refreshes into 1 frame (one for the signal and one for each of two possible masks). There were three different temporal presentation modes: 1) Stimuli were presented for a single frame, i.e. 3 refreshes, nominally for $19.7ms$. 2) Stimuli were presented for 4×3 frames, nominally $79ms$. 3) Stimuli were presented with the contrast of all components following a Hanning window of $1497ms$ total duration. All reported contrasts are the peak contrast at the center of the time interval. To extract thresholds from the data I fitted the data using psignifit 4 with the standard prior set based on the tested stimulus range (Schütt, Harmeling, Macke, & Wichmann, 2016). Error-bars represent 95% credible intervals.

I also present data from the Modelfest dataset (Watson & Ahumada, 2005) and a natural image masking database (Alam et al., 2014) here. The Modelfest dataset consists of contrast detection thresholds for 43 different 256×256 pixel targets presented at 120 pixels per degree. Target contrast was temporally modulated by a Gaussian envelope with a standard deviation of $125\,ms$. The natural image masking database consists of the detection thresholds for $3.7\frac{cyc}{deg}$ log-gabor filtered noise targets masked by 1080 natural image patches taken from 30 black and white digital photographs. Thresholds were measured using a spatial three alternative force choice task. Three stimuli were presented simultaneously and subjects had 5 seconds to indicate which stimulus contained the noise Gabor target overlayed over the natural image patch. Further technical details for these datasets are provided in the original studies.

2.3 RESULTS

2.3.1 *Classical psychopysical results*

I first test my model on classical psychophysical experiments. These experiments were specifically designed to test hypotheses about early spatial visual processing. To achieve this, the stimuli are composed of sinusoidal gratings intended to activate the spatial frequency and orientation channels as specifically as possible. I start with the sensitivity of single channels and continue with masking experiments, which test how well activation of additional channels mask the signals.

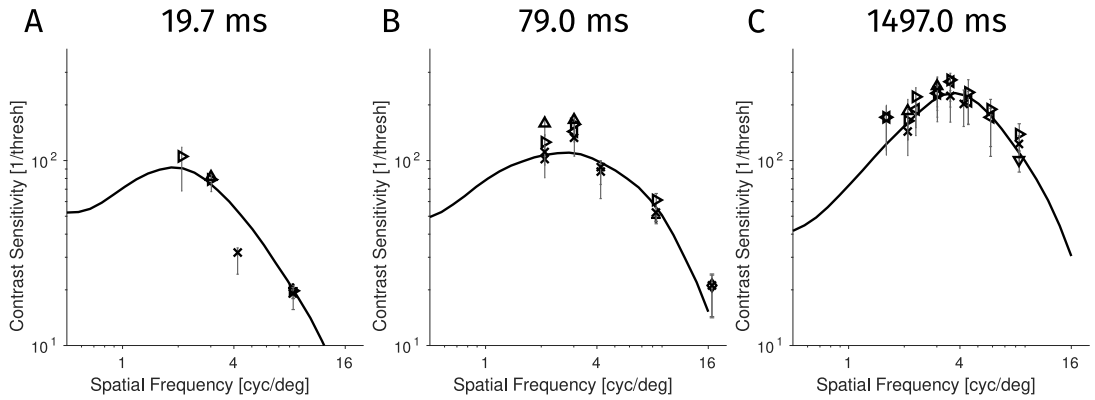

Figure 7: Results for the contrast detection data for different presentation times. Different symbols represent the measured data from different observers. Each observer has their own fixed symbol accross all figures. Errorbars represent 95 % credible intervals from a Bayesian analysis of individual psychometric functions. The continous line represents the prediction of the model. **A,B**: 19.7 and 79 ms (3 and 12 frames) presentation time with hard on and offsets. **C**: Contrast Hanning windowed in time with a total presentation time of 1497 ms

Contrast detection

I present detection data for three different temporal presentation modes, roughly 20ms and 80ms with hard on and offsets and contrast changing according to a 1.5 second long Hanning window/raised cosine window.

The data are presented in the form of *contrast sensitivity functions* (CSFs) in Figure 7. The contrast sensitivity functions show the typical bandpass shape for long presentation times and the more low-pass shape for the short presentation times.

The model reproduces the contrast sensitivity functions closely. This is not surprising as I fitted a weighting for the spatial frequencies in the preprocessing for each presentation time.

ModelFest

Next I evaluate the model against the ModelFest database, incorporating detection performance for 43 different patterns measured with many observers in different labs.

The evaluation of my model for these data are displayed in Figure 8. First I ran the model with a new contrast sensitivity function and the parameters fitted for the adjacent presentation times. With these parameters I already obtained promising fits displayed as the grey lines in Figure 8, which fitted almost all patterns in the data. The main error seems to be a constant offset, which could probably be corrected by adjusting the initial weighting filter. Using parameters fitted to the data, I obtain an even slightly better fit to the data plotted as the black line in Figure 8.

The clearly largest deviation from the data for all parameter settings is the Gaussian blob (stimulus #26). This very low spatial frequency target is strongly affected by the initial luminance normalization. Consequently I believe that this represents a problem of the overly simplistic preprocessing, which ignores stimulation before and after the stimulus, which sets the adaptation level differently from the mean of the image presented.

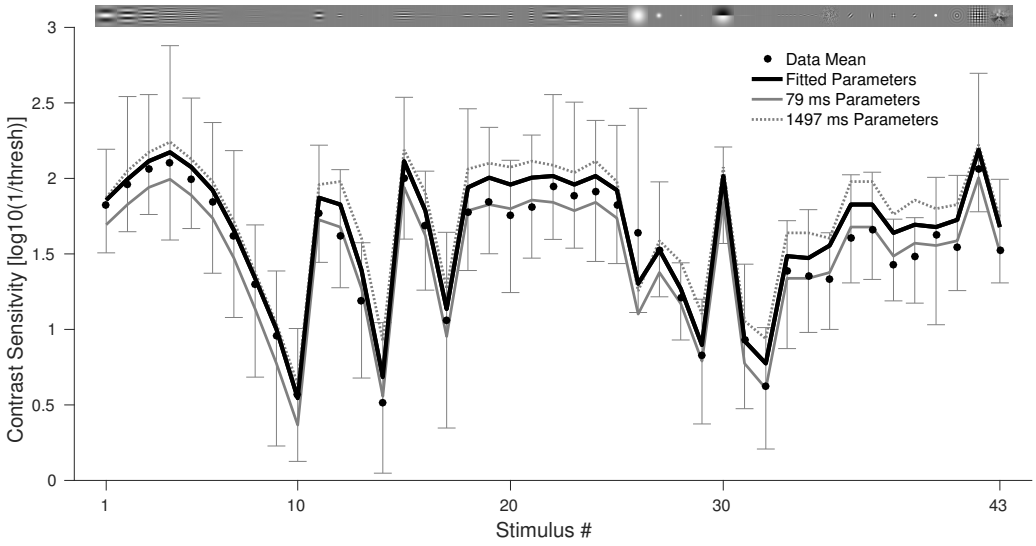

Figure 8: Results for the ModelFest dataset. Here I plot (log-) contrast sensitivity for the 43 different stimuli ordered along the x-axis. The dots represent the average measured threshold, with errorbars representing the range of measured thresholds. The lines represent the predictions of my model using different parameters. Above the plot tiny full contrast images of the stimuli are displayed.

Contrast discrimination

The next type of data I compare my model to is contrast discrimination data, which originally motivated the nonlinearity (Foley & Legge, 1981; Legge & Foley, 1980). Here the task is to report which of two presented gratings has the higher contrast, i.e. to discriminate gratings, that differ only in contrast.

I start by investigating only the $78.8\,ms$ presentation data presented in Figure 9. At all spatial frequencies the thresholds for discrimination follow the classically observed dipper shape (Foley & Legge, 1981; Legge & Foley, 1980). All curves first decrease such that at low pedestal contrasts, contrast discrimination is easier than detection (Nachmias & Sansbury, 1974). At higher contrasts, discrimination thresholds lie roughly on a straight line in the log-log plot indicating a power law for the contrast discrimination threshold.

The model reproduces the contrast discrimination curves quite well for all spatial frequencies. Also the slopes of the psychometric functions seem to be captured by the model, since I fit thresholds at different performance levels. Especially the shallower psychometric functions in the dipper reported by Bird et al. (2002) are reproduced.

Next, one can investigate how contrast discrimination performance varies with presentation time. For the $8.37\frac{cyc}{deg}$ target contrast discrimination data was also available at the two other presentation times of $19.7\,ms$ and the $1497\,ms$ hanning window.

These data with model fits are plotted in Figure 10. In each panel I show the data measured with given presentation time together with three different fits. All of these fits use the contrast sensitivity filter fitted for the correct presentation time, but normalization parameters fitted to the three presentation times. The model can reproduce the data for each presentation time. However, the different presentation times require different parameters, since the curves simulated from a single parameter set do not capture the data adequately. Especially the width of the dip and its position relative to the detection threshold differ between presentation times.

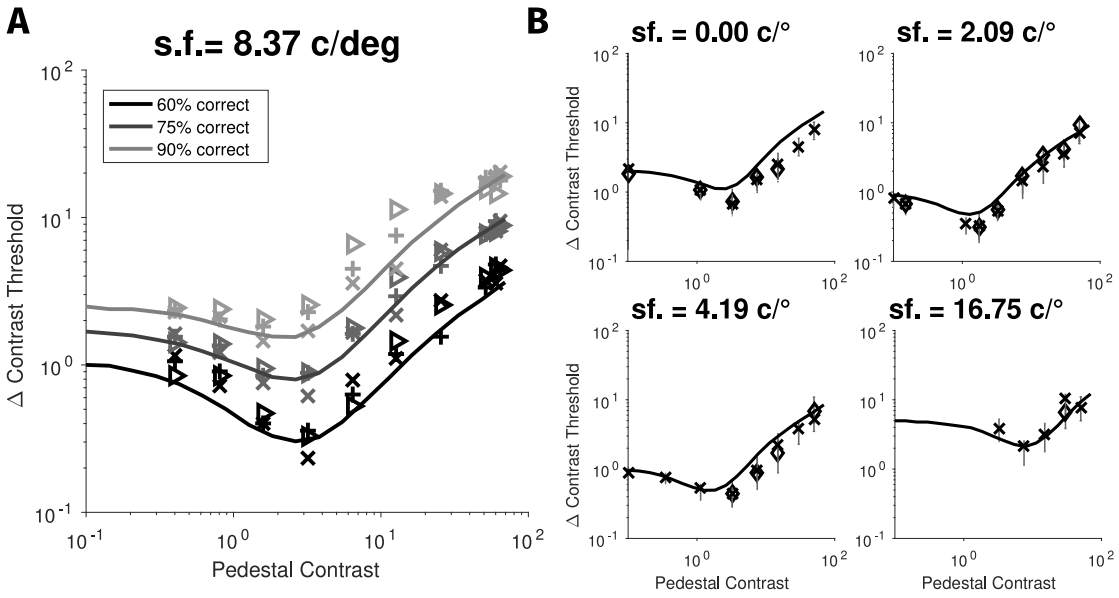

Figure 9: Results for contrast discrimination data. All data were collected with $79ms$ presentation time with hard on and offsets. **A**: Data for $8.37\frac{cyc}{deg}$, the frequency for which most data was available. The different grey values indicate different percent correct to be reached to define the threshold. This illustrates the change in the slope of the psychometric function over the range of contrasts. Specifically it is shallower in the dip and steepest for detection. **B**: Results for different spatial frequencies. Here only the data for the 75% contrast are shown. $0.00\frac{cyc}{deg}$ indicates discrimnation in the brightness of a blob. All other conventions are as in Figure 7.

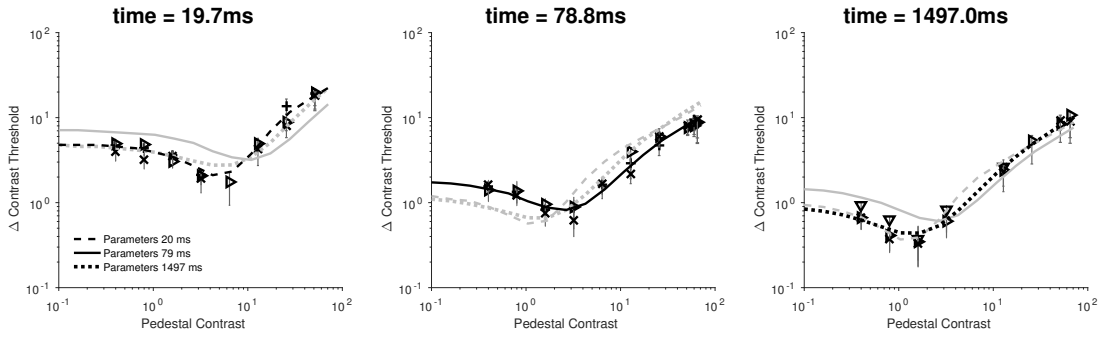

Figure 10: Results for contrast discrimination datafor different presentation times. Each panel shows the contrast discrimination data for the $8.37\frac{cyc}{deg}$ for one presentation time. Again different symbols show the 75% threshold from different observers with 95% credible intervals. The lines represent the predictions from three different sets of parameters. In each panel the prediction with parameters fit to the displayed data is highlighted in black. All other conventions are as in Figure 9.

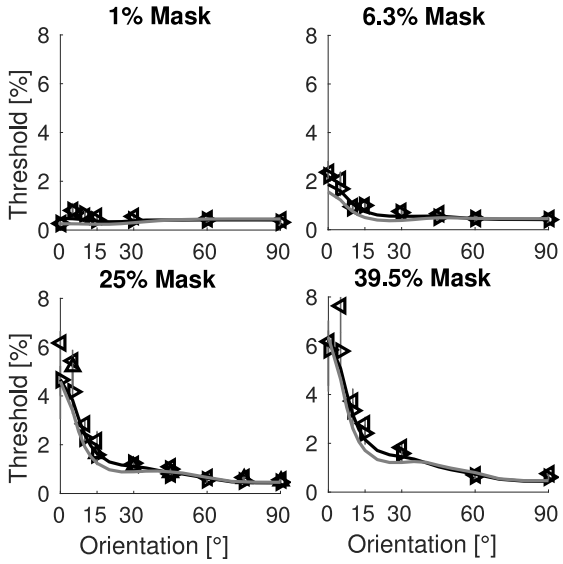

Figure 11: Results for oblique masking experiments, spatial frequency for both signal and mask was $3\,\frac{cyc}{deg}$. As in previous figures, symbols represent data and lines the predictions of my model. The black line uses parameters specifically fit to the oblique masking data, the grey line is the prediction using the parameters estimated using all data at the long presentation time of $1497ms$

Oblique masking

Next I compare my model to oblique masking data, which represent the psychophysical reason for replacing the channel wise nonlinearity with normalization across channels (Foley, 1994). Here the task is to detect the presence of a horizontal grating, while all observation intervals contain an additional "oblique mask", i.e. another grating of the same spatial frequency and spatial envelope, but with a different orientation. All oblique masking experiments were performed with the $1497ms$ presentation time and $3° \times 3°$, $3\frac{cyc}{deg}$ targets.

Results of these experiments are presented in Figure 11. While the masking effect of nearby orientations is slightly underestimated by the model the overall fit of the model to the data is good.

Plaid masking

The next type of data I compare my model to is plaid masking data. Here, the task is the same as for oblique masking, but the one oblique mask is now replaced with two masks rotated away in opposite directions from the signal orientation, which are together called a plaid.

Results of these experiments are displayed in Figure 12. Characteristic for these experiments is that at relatively high contrast (here 25%) plaids 30-45° and even further away from the signal orientation substantially mask the signal, while each of the two gratings composing the plaid alone hardly mask the signal. Thus, the two gratings masking capabilities combine strongly super-additively. To show this super-additivity I replotted the oblique masking data in the figure.

My model fails to replicate the super additive masking effect of plaid masks, as most probably all other spatial vision models based on the multi-resolution theory do (Derring-

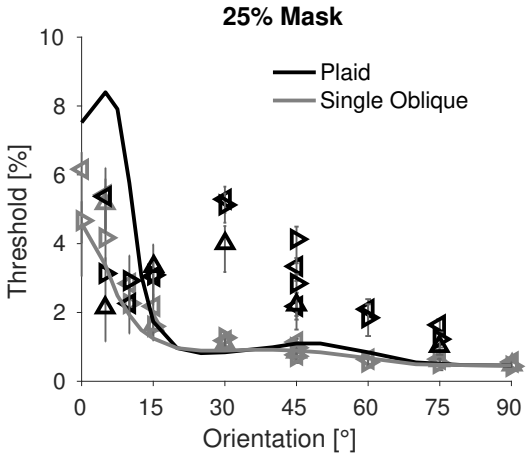

Figure 12: Results for Plaid masking experiments, here only for the 25% contrast mask. The plaid data and model predictions are plotted as the black symbols and line again. Addtionally I replot the data and prediction for a single oblique mask from Figure 11 in grey.

ton & Henning, 1989). A clearly favoured explanation of this effect has not yet emerged although it is strong and reliable. For some weaker forms of plaid masking where the signal and mask are separated in spatial frequency linear summations over channels can explain plaid masking (Holmes & Meese, 2004). For the effects of plaids of the same spatial frequency only speculations exist though. One is that plaid masking is a perceptual effect created because observers frequently perceive high contrast plaids as "checkerboards" oriented between the orientations of the plaid components (Georgeson & Meese, 1997). A different one is that the recurrent dynamics of V1 might create activity at orientations different from the signal orientations, especially at the orientation between the two plaid components (Carandini & Ringach, 1997). However, neither of these suggestions can be easily incorporated into the kind of model I propose here.

2.3.2 *Natural scene masking database*

To include some evaluation of my model on more natural stimuli than gratings, I evaluate my model on a natural image masking database (Alam et al., 2014). The database consists of the detection thresholds for log-gabor filtered noise targets masked by 1080 natural image patches taken from 30 black and white digital photographs.

To apply my model, I used a single exemplar of the noise, which accompanies the database and calculated its detectability on the different patches imitating the conditions the subjects saw in the experiment as closely as possible. As subjects were allowed to move their eyes and my model cuts out a rather small foveal area, I simulated not only a fixation at the exact center of the patch and signal, but also at the 8 points moved 0.5° up and down and/or left and right from the center. Following the overarching theme of optimality, I display the lowest of the 9 thresholds obtained this way. For the parameters, I chose the parameters for the long, 1.5 second Hanning window as the natural image patches were displayed for an even longer time of 5 seconds.

To convert the images to luminance values, I used the formula provided with the database, although it returns values smaller than the minimum luminance of the monitor reported in the paper. Thus, the data for dark patches seems to be unreliable. Also, the original paper excluded patches with low average luminance. Consequently, I follow the

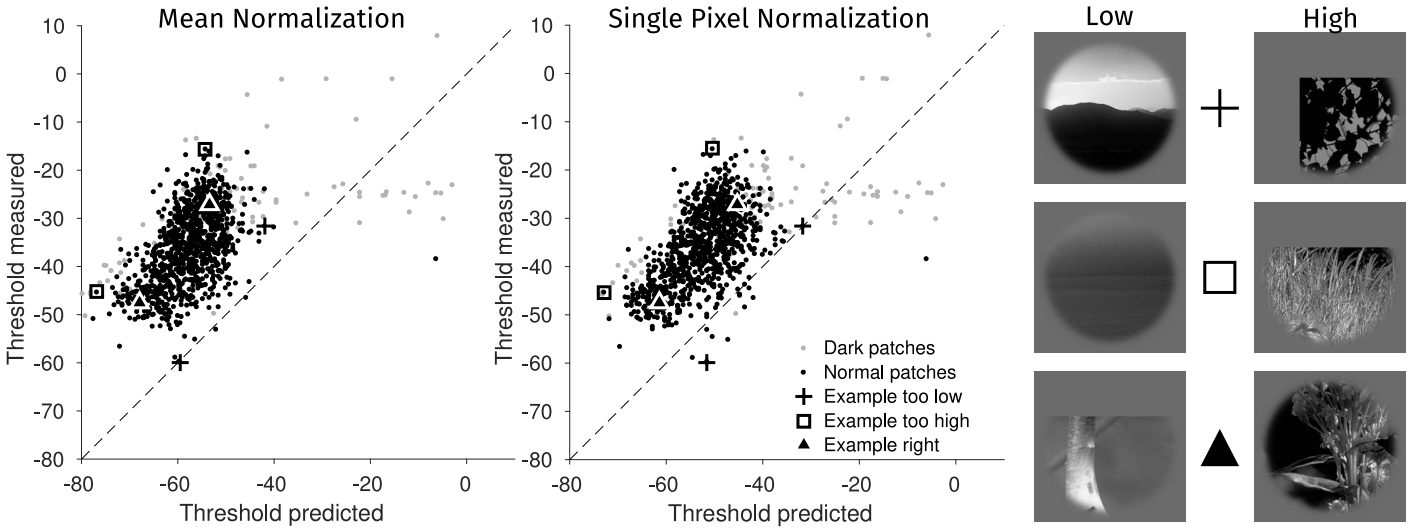

Figure 13: Results for the natural image masking database. First I plot the measured thresholds against the predictions of my model setting the spatial extent of the normalization pool either to the whole image or to a single pixel. Patches darker than $4\frac{cd}{m^2}$ are plotted in grey, all others in black. Additionally, I marked one low and one high threshold example patch each, where the measured threshold was higher, lower or roughly equal to the prediction.

lead of the original paper and exclude patches with an average nominal luminance below $4\frac{cd}{m^2}$ from further analysis. These excluded patches are still displayed in Figure 13 as grey dots.

The results of my model are displayed in Figure 13. I find that the model generally overestimates the sensitivity of observers on the natural image stimuli, but produces thresholds highly correlated to the measured ones and thus seems to represent a sensible upper bound on these data. Models designed and adjusted specifically to fit this database can produce higher correlations with the data (Alam et al., 2015, 2014). Nonetheless, for generalization from grating based experiments, the predictions seem to be quite accurate. Also, the model errs in the explainable direction. It seems plausible that highly trained observers perform better on simple grating stimuli without any random variation than less trained observers on natural image patches whose exact properties they were not extensively familiar with.

Surprisingly, I find that the single pixel normalization scheme, which was problematic for predicting the classical grating data, yields a higher correlation to human thresholds ($r = 0.5801$) than the mean normalization scheme ($r = 0.5196$), which was better at predicting the grating data. Tentatively, I assume that there is a local normalization scheme of medium size, which still fits the grating data and produces an equally good prediction as the local normalization.

One possible explanation for why my model predicts too low thresholds for the natural image stimuli might be that subjects are worse at decoding the noise signals on the natural image masks than they are in the simpler classical grating experiments. In Figure 14 I show one specific weaker decoder. Namely, it weights any difference only by its sign instead of its signal to noise ratio. This results in a decoder that simply adds all image differences, but ignores how well the specific channel differentiates the two images. This scheme is equivalent to taking the Minkowski-1-norm of the difference between the images drawing the connection to earlier models. Clearly such a simpler, worse decoder moves the

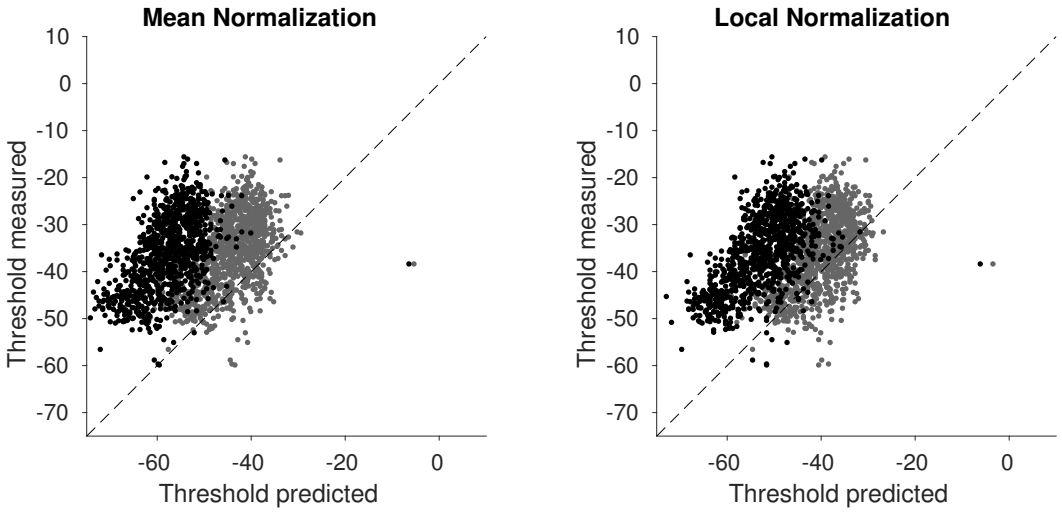

Figure 14: Using a weaker decoder for predicting the early natural image masking database. The grey dots represent the predictions from the model when differences between the two images are summed disregarding their signal strength. The black symbols reproduce for the optimal decoder from Figure 13.

predictions much closer to the measurements. However, I do not claim that this specific decoder mimics human behaviour, as many other bad decoders would certainly increase the predicted thresholds equally. Nonetheless this illustrates the point that a realistic but suboptimal decoding could explain the weaker performance of subjects in this natural image masking task.

2.3.3 *Different parameter sets*

To further investigate the models' internal processing, we shall have a look at how the parameters needed to be changed to fit the different presentation times and data types. As described in detail in Appendix A.1, I first fit the longest presentation time for which the oblique masking data was available to fix the orientation bandwidth of the normalization pool and then fit the parameters of the final normalization for the different presentation times and for ModelFest.

The parameter fits are given in Table 1. First, note that the linear contribution to the noise N_f is 0 for all datasets. I set this because I noticed that the exponent q can compensate for vastly different N_f such that all of them explain the data equally well (see Appendix A.1). Additionally, there is a presentation time dependent scaling of the input in the model. Thus, the constant C cannot be compared directly across presentation times. Consequently, only the exponents p, q and possibly the noise strength N_C can be compared between presentation times. Furthermore, the parameters I fitted for ModelFest depend on the data augmentation used to achieve a good fit from the thresholds only and the oblique masking data were fit to a considerably different kind of data. Thus, I only discuss the parameter sets fit to the contrast discrimination data at the three presentation times.

For these three presentation times q—which regulates the high contrast behaviour of the model—changes little with the presentation time. This observation corresponds to the empirical statement that the power law behaviour at high contrasts has a similar log-log slope for all presentation times. The exponent p changes such that longer presentation

Table 1: Parameter values used for the different experiments. The bold values were fit for the data in the experiment, the others were kept at the values estimated from the 1497ms presentation time data, as most of the available oblique masking data was available at that time.

parameter	meaning	19 ms	79 ms	1497 ms	oblique	ModelFest
N_c	const. noise Var	**1.4389**	**0.6450**	**0.4763**	**0.4235**	**0.0070**
N_f	noise factor	0*	0*	0*	0*	0*
C	NL constant	**0.0031**	**0.0046**	**0.0027**	**0.0014**	**0.0147**
p	NL exponent	**2.7996**	**2.0253**	**1.8667**	**1.3732**	**1.2090**
q	difference exponents	**0.3767**	**0.3676**	**0.3032**	**0.3755**	0.3032
ω_θ	norm. pool orient.	0.2008	0.2008	0.2008	**0.2008**	0.2008
σ_θ	filter std. orient.	0.2965	0.2965	0.2965	0.2965	0.2965
ω_f	norm. pool freq	1	1	1	1	1
σ_f	filter std. freq	0.5945	0.5945	0.5945	0.5945	0.5945
$\omega_x = \omega_y$	norm. pool space	—	—	—	—	—

times require a lower exponent. This fits the empirical observation of a less pronounced dip at longer presentation times (see Fig. 10).

Additionally, the noise variance N_C decreases with presentation time fitting the absolute decrease in thresholds for longer presentation times. This could be interpreted as averaging away noise over time. However, caused by the different scaling of contrast applied before the decomposition and the different C it is not entirely clear whether this conclusion should be taken seriously based on these data.

2.3.4 *Analysis of the models representation*

Additional to the theories developed based on psychophysical or neural measurements, researchers developed normative theories to characterize what the information extracted from natural stimulation for animals or humans should be. My model was not designed to maximize coding efficiency or to fit natural stimuli. Thus it is interesting to have a more detailed look what responses to natural stimuli look like and which normative principles my model follows.

As a first qualitative analysis on the model output, I looked at the responses my model produces to natural images. Simply summing the responses from all channels I found that my model indeed highlights edges. This fits the earliest accounts of the responses of primary visual cortex neurons (Hubel & Wiesel, 1968). As an example, I show the summed response for the example photograph of the Tübingen town hall in Figure 15 A. To allow a better display I show the square root of the sum. Note also that the town hall is easily recognisable from this representation.

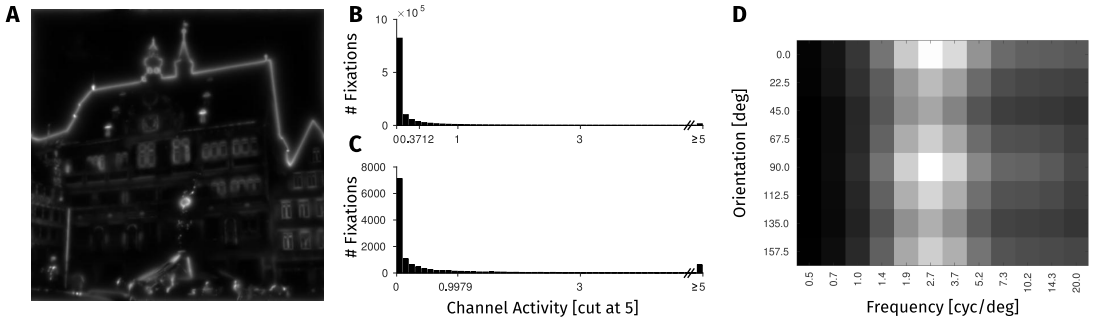

Figure 15: Some information on the output of the model. **A:** Square root of the sum of the outputs of all channels for the example photograph of the town hall of Tübingen as an example unrelated to panels B,C and D. The output of my model highlights edges. **B:** Histogram of all channel activities over fixation locations in natural images. As the highest channel activity I observe is 194, I cut the histogram at 5 to make some distribution visible. The activity distribution is extremely skewed, i.e. my model produces a sparse code. **C:** As in B, but only for the most active channel (vertical with $2.7\frac{cyc}{deg}$ peak sensitivity) to show that each channel is sparsely active. **D:** Mean activation over all fixation locations.

2.3.5 *Sparseness*

To get some more quantitative information about the typical responses of my model, I analysed the responses of my model to some natural images, for which eye movement data are available from an earlier study (Engbert et al., 2015). In this study 35 observers explored 15 natural scenes and 15 photographs of texture surfaces for 10 seconds each to memorize them. During this experiment they produced 24582 fixations. At each of these fixations I extracted the activity at the fixated pixel from an image I had processed by the model as a whole without the foveal window. This might give us some hint what the internal representation in my model looks like for natural foveal stimulation of human observers.

First, I looked at the range of activations observed and found an extremely skewed distribution (see Fig. 15 B): Maximal activations were almost 200, while 98.7% of the channel activities observed were smaller than 5. This effect is caused by skewed distributions in each channel. To illustrate this I show the activity histogram of the most active channel in Figure 15 C. Even this most active channel is rarely active. These observations fit well with theoretical arguments for using a sparse code (Olshausen & Field, 1996) and physiological observations showing sparse neuronal responses (Buzsáki & Mizuseki, 2014).

To quantify the sparsity of the model responses, I used the formula developed first by Rolls and Tovee (1995) and refined and applied to primate primary visual cortex by Vinje and Gallant (2000):

$$S = 1 - \frac{\left(\frac{1}{n}\sum_{i=1}^{n} r_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \frac{1}{1-\frac{1}{n}} \quad . \tag{8}$$

S measures the proportion of the sum of squares explained by the mean response and subtracts it from 1. After dividing by $1-\frac{1}{n}$ this yields a measure which conveniently scales from 0 to 1 from a constant response to a perfectly sparse response, which reacts exactly to one stimulus and is 0 for all others. Applying this formula to the model responses I follow Froudarakis et al. (2014) in separating *population sparseness*, i.e. whether the

population response to a stimulus is sparse, from *lifetime sparseness*, i.e. whether an individual channel is sparsely active over the presentation of all stimuli.

For population sparseness, I find an average value of 33.86% for the raw decomposition and 52.31% for the normalized responses, which is more sparse than average neuronal populations in mouse V1 (mean = 0.26, max $\leq$ 0.6) as measured by Froudarakis et al. (2014), but within the range observed. Due to the small numbers of simultaneously recorded neurons in typical primate recordings we lack data to compare my model to for monkey primary visual cortex.

Investigating lifetime sparseness, I find high values for the sparseness of the channels as displayed in Figure 16. On average the channels after the raw decomposition have $S = 55.07\%$, which increases to an even higher S of 73.85% after normalization. These are both much higher than the lifetime sparseness measured in mouse V1 by Froudarakis et al. (2014), which was 35% on average.

Furthermore, my model also reproduces the observation that natural stimulation—viewing natural images—elicits a sparser code. Patches extracted around fixated locations yield higher lifetime sparseness in high spatial frequency channels than control patches, which I extracted at the measured fixation locations, but from different images from the stimulus set (see Figure 16).

I also computed the average activations produced by the channels in my model. The results are displayed in Figure 15 D. After the normalization the fall off for higher spatial frequencies inherent in natural images (Field, 1987) is not observed any more. In contrast, the higher content for the cardinal axes (0 and 90° in my notation) persists after the normalization (Furmanski & Engel, 2000; B. Li, Peterson, & Freeman, 2003). This activation pattern qualitatively fits reasonably well to the distribution of neurons in primary visual cortex, fitting the idea that the distribution of neuronal preferences reflects the distribution of activations produced by natural stimulation (Field, 1987; Laughlin, 1983).

2.3.6 *Optimized stimuli*

One additional benefit of (successful) image computable models of human vision is that they should allow the generation of image modifications leading to minimal and/or maximal perceptual differences, exploiting the idea of maximally differentiating (MAD) stimuli (Wang & Simoncelli, 2008). In the following I illustrate the viability of MAD applied to my image computable spatial vision model, comparing changes in the model responses to the default and simple root mean squared error (RMSE) metric.

For the illustration I optimized the images to be as easy or as hard to differentiate from the image of the Tübingen town hall as possible with a given RMSE after conversion to luminance and application of the foveal window. The exact optimization scheme is described in detail in Appendix A.3.

In Figure 17 I show three images with equal RMSE from the original Tübingen town hall example image: One with minimized differentiability, one with simple Gaussian noise and one with maximized differentiability. The optimization clearly produced stimuli which are predicted to be considerably more or less differentiable from the original image but all have the same RMSE.

In the image with maximized differentiability we can observe two aspects of the model: First, a single, local signal is predicted to be more easily detectable. Second, the optimized signal is similar to the filter shape of a single channel of medium spatial frequency where contrast sensitivity is highest.

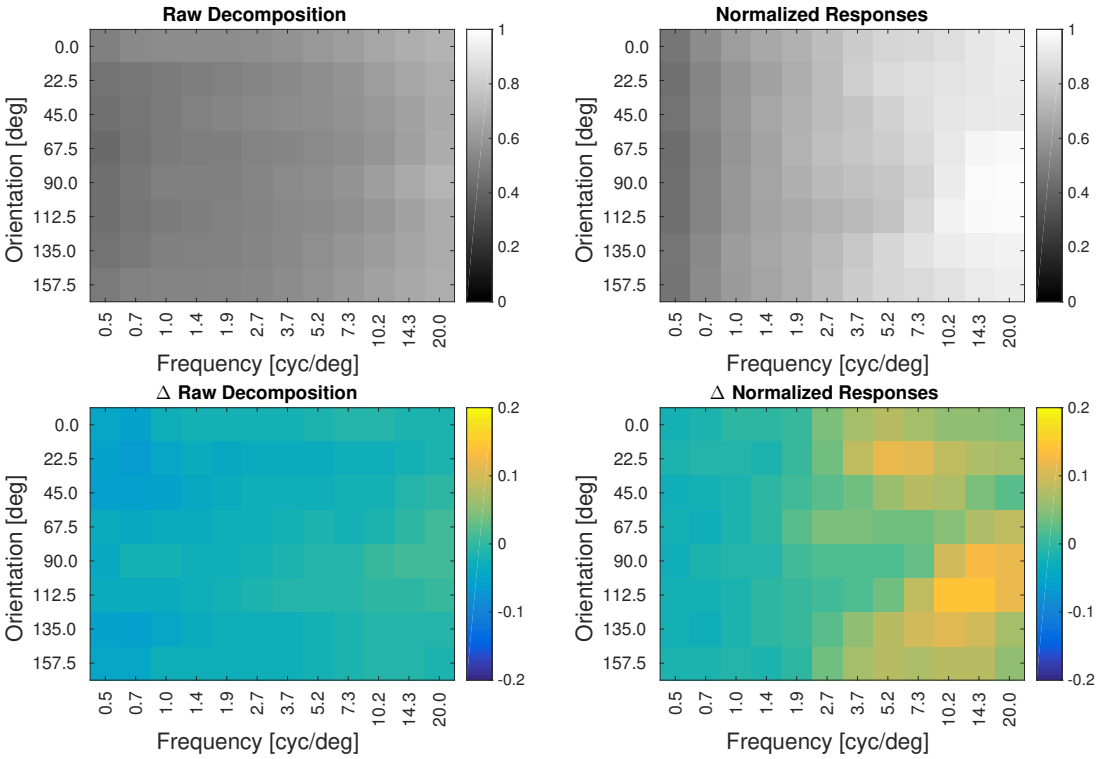

Figure 16: Lifetime sparseness for the different spatial frequency and orientation channels. Left shows the sparseness of the linear filter responses (before nonlinearities and normalization). Right shows the sparseness of the final responses. In the top row I show the sparseness of activities at fixated locations. In the lower row I show the difference between the sparseness at fixated locations and the sparseness at non-fixated control locations.

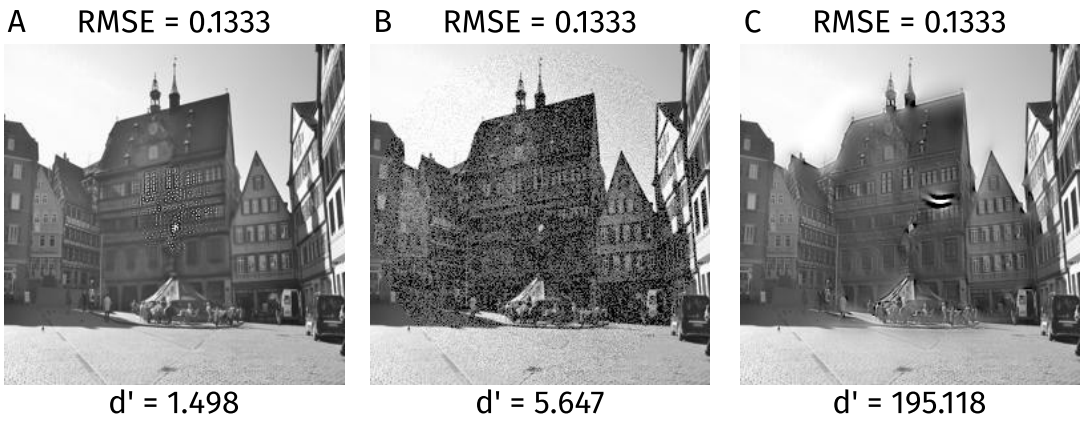

Figure 17: Stimuli with optimized differentiablity from the original Tübingen town hall image, with a given RMSE in the windowed contrast image. Luminance images are displayed assuming a gamma of 2.2. In the model these images were simulated to cover $2 \times 2°$ of visual angle. **A**: Minimized differentiability, **B**: Gaussian over the area withing the foveal window, **C**: Maximized differentiability

In the image with minimized differentiability the RMSE is realized as a high frequency non-oriented and distributed noise on the image. This indeed becomes practically invisible when viewed such that the image covers the $2 \times 2°$ simulated in the model (around a 1.4 m distance if you received this thesis on A4 paper, such that the images are 5×5 cm).

These generated stimuli demonstrate that my model is capable of producing predictions for supra-threshold stimuli and their differences, which are interpretable and testable. This makes my model potentially applicable for image quality assessment, and, as discussed below, allows more thorough tests of the model to be performed.

2.4 DISCUSSION

I describe an image-computable model of spatial vision. When applied to classical psychophysical results it is consistent with the broad range of contrast detection, discrimination and oblique (orientation) masking data fitted by earlier, more abstractly implemented, non-image-computable models. In addition, I tested the model on the ModelFest dataset on which it also performs well. Alas, my model—like all previous models—fails to account for human plaid masking data.

Whilst developing my model, I uncovered two crucial ingredients for a successful image-computable spatial vision model: First, when including nonlinear interactions between channels after the decomposition, strong oscillations in the response are observed unless I sample the spatial frequency and orientations axes more densely than required from signal processing considerations. In the human visual system this appears to be solved by not having discrete channels like in engineered subband transforms, but by having a continuous distribution of cells covering the relevant spatial frequencies and orientations.

Second, different temporal presentation modes require—systematically—different parameters of the model: shorter presentation times require higher exponents for both the signal and the normalization pool, yielding stronger nonlinearities for short presentation times. This finding confirms an earlier conjecture by Wichmann (1999), based on much simpler models, and might explain differences in estimated exponents between different labs and studies. For the parameter q, the difference in the exponent between numerator and denominator—I find little dependence on the presentation time once I assume

only a constant noise. Finally, variance of the constant noise decreases with presentation time. All these changes in parameters are consistent with the following picture: Channels show an onset response with stronger nonlinearity, followed by a less nonlinear sustained response. Over time human observers appear to be able to average some of the noise.

When I applied my model to the natural image masking database by Alam et al. (2014), I found that my model predicted the data reasonably well, but almost always predicted lower thresholds than observed in their experiment. Potential reasons for the discrepancy include the following: First, the optimal decoder knows both signal and mask exactly, which is unlikely to be true for human observers with either stochastic or hitherto unseen natural images as masks. Thus the overestimation of performance of my model may in part be due to using a too knowledgeable decoder. Second, my model is solely fit to data from very experienced psychophysical observers, and I do not know how experienced the observers in the natural image masking study were (c.f. Jäkel & Wichmann, 2006).

To investigate whether my model conforms to normative notions derived from efficient coding, I analysed its response to natural images at positions fixated by human observers. I found that my model produces sparse responses as expected from theoretical considerations. Furthermore, average responses still contain a bias for cardinal orientations as observed in natural images, but the 1 over f decline over spatial frequency associated with natural images is obliterated by normalization.

Finally I created MAD stimuli to compare the model to the RMSE. These illustrate the behaviour of my model. Additionally, such stimuli might be used to psychophysically test my model in the future, which is the intended purpose of MAD stimuli (Wang & Simoncelli, 2008). Especially once one wants to test different, more complex models against each other, analyses like this are invaluable.

2.4.1 *Comparison to earlier models*

As I specifically designed my model to be an image-computable version of the standard spatial vision model, it naturally shares many properties with earlier models and implementations.

The model by Foley (1994) first introduced the spatial frequency and orientation channel decomposition followed by divisive normalization, which is at the heart of my model. However, Foley implemented decoding as a Minkowski norm of the difference between responses instead of explicitly modelling noise and optimal factorial decoding as I do here. Another model using the simpler Minkowski norm decoding scheme is the model by Itti et al. (2000). This model is also an important precursor of my model, as it showed that different tasks like spatial frequency and orientation discrimination could be explained by a *single* model of the style I use here. Finally, the most closely related abstract, non-image-computable model is the model by Goris et al. (2013). The remaining conceptual differences of my model to the Goris model are, on the one hand, that I did not include noise correlations or adaptation present in the Goris model, but, on the other hand, I added the spatial extend of the normalization pool, orientation, etc., to move my model from 1D to 2D.

Of the few image based spatial vision models the two most closely related ones to ours are the models by Teo and Heeger (1994) and by Watson and Solomon (1997), which both implement a spatial frequency decomposition and divisive normalization. However, they use the simplistic Minkowski norm decoding and were implemented with the technology of their time, which made diverse compromises for speed necessary. For example, the Watson and Solomon (1997) model represented only three spatial frequency channels of

which one always hit the spatial frequency of the target. Also both models were compared to a rather small range of data and some views and questions like natural scene statistics and optimal coding were not yet discussed at the time of these models.

Most other image-computable models of spatial vision do not aim to mimic the internal processes involved in spatial visual processing, but simply optimize prediction with computationally less demanding processes. Especially the most modern models of this kind (Bradley et al., 2014) predict human performance quite well and can even include peripheral limitations. However, these models are designed for different purposes than my model providing no output similar to the output of the first steps of the human visual system and allow no direct tests of hypotheses about the early visual processing either.

2.4.2 *Potentially controversial details*

Phase invariance

My model provides phase invariant output, which represents the information perfect complex cells would convey. This is computationally efficient and provides all information necessary for the psychophysical tasks I model. Additionally, this nicely fits with other psychophysical data which explicitly shows phase independence for the detection of multiple sufficiently separate components (Graham & Nachmias, 1971) and that phase perception can be explained based on detection of local contrast changes (Badcock, 1984, 1988). However, neuronal data show that the distinction between simple and complex cells is gradual and both types express some sensitivity to relative phase (Mechler, Reich, & Victor, 2002). Furthermore humans show more dependence on phase information for object recognition than predicted from contrast reduction caused by phase noise (Wichmann, Braun, & Gegenfurtner, 2006). Consequently a more complete model might add decoding from phase dependent output to mimic simple cells, or even include both simple and complex cells.

Tuning and complexity of the normalization pool

The spatial vision community is divided whether the normalization pool is orientation specific. In my purely divisive normalisation implementation—without a subtractive normalization—orientation specific normalization is required to be consistent with my data; the same is true for the model by Itti et al. (2000). The models by Teo and Heeger (1994) and by Foley (1994) argue for an orientation unspecific normalization, in line with neurophysiology (Heeger, 1992).

In our data and the data of Itti et al. (2000), orthogonal gratings barely mask each other, even at high mask contrasts—thus a non-tuned divisive normalization does not fit such data. In the data by Foley (1994), however, orthogonal gratings mask the signal grating. Similarly physiologists sometimes find that orthogonal gratings considerable attenuate neuronal responses (Heeger, 1992)—cross-orientation inhibition. However, at least the suppressive surround is sometimes found to be tuned (Cavanaugh, Bair, & Movshon, 2002b). One possible explanation for this discrepancy *in the data* is the temporal presentation of the stimuli during the experiments. Our data and the Itti et al. (2000) data were collected using static gratings presented for an extended period of time, while the data of Foley (1994) and Foley and Boynton (1994) were collected using very short presentation times. Thus the normalization pool may initially be broadly tuned, but narrows during prolonged presentation.

Furthermore, the normalization I implemented does not cover all interactions reported between channels. There are well known facilitatory effects of collinear flankers (Polat & Sagi, 1993). Most commonly these are interpreted as facilitatory effects between channels, but alternatively these could be explained by collector units further up in the hierarchy of visual processing (Solomon & Morgan, 2000). Similar ideas were also proposed to explain the unexpectedly strong masking produced by amplitude modulated gratings at their modulation frequency (Henning et al., 1975). Such explanation based on further processing of the filter responses are compatible with my model being a correct model of the first transformations in spatial vision. If the interpretation as facilitatory effects in the earliest representation is correct however, it should be included in future spatial vision models.

High contrast signals

Another aspect differing between models is how they treat high contrast signals. In my model I implement a higher numerator exponent in the normalization, which yields non-saturating responses in the individual channels as in the model by Watson and Solomon (1997). The alternative approach followed by Teo and Heeger (1994) is to simulate multiple types of channels covering different contrast ranges (in their case four). This second approach models the responses in closer agreement to neuronal data, as neurons undeniably saturate. From a psychophysical perspective this seems to add little however, as channels differing only in their absolute sensitivity cannot be targeted specifically by any stimuli and are thus modelled quite adequately as a single channel. Only if the cells or channels for higher contrasts had different tuning curves or interactions with other channels than the low contrast ones it were necessary to separate them. Consequently, I interpret the V1 neurons for different contrast levels as the neuronal implementation of a single channel using multiple neurons to avoid saturation.

Decoding stage

For decoding I follow modern abstract models like the Goris et al. (2013) model and explicitly model the noise on individual channels and propose optimal or near optimal read out of the channel responses in a Bayesian sense (c.f. Beck et al., 2008; Ma, Shen, Dziugaite, & van den Berg, 2015). The idea that observers in basic psychophysical tasks are (only) limited by an internal noise source has recently been challenged. Beck, Ma, Pitkow, Latham, and Pouget (2012) instead propose that performance is limited by imperfections of the read out mechanism. For explaining the systematic discrepancy between my model and natural image database data by Alam et al. (2014), I follow this interpretation. It appears that (highly experienced) observers during simple contrast detection and discrimination experiments were more sensitive than subjects producing the natural image masking data. I suggest that this might be caused by better decoding rather than more available information, similar to the suggestion that perceptual learning improves decoding rather than the original representation (Diaz, Queirazza, & Philiastides, 2017). In my model the decoding is optimal for the classical grating experiments, as these experiments are set up to make decoding as easy as possible for humans.

Variance and type of internal noise

The noise model used in early spatial vision models has always been a matter of discussion, partly because the psychophysical data collected during classical detection and discrimi-

nation tasks appear not to constrain the standard model sufficiently. Even fundamental questions as whether the noise variance changes with signal strength were not finally answered by psychophysics yet, although some attempts were made (Georgeson & Meese, 2006; Kingdom, 2016; Kontsevich, Chen, & Tyler, 2002; Wichmann, 1999). Based on the maximum-likelihood estimation I cannot, unfortunately, answer the question whether the noise grows with the signal or not. My model can explain the data with constant noise equally well as with noise variance growing linearly with the signal. The underlying reason for this is that changing the q-parameter can compensate for a growing noise. In terms of the neural implementation this corresponds to the statement that adding more neurons tuned to high contrasts can compensate for neurons being noisier when responding strongly. This insight explains why I cannot differentiate how the noise should change with increasing contrast based on psychophysics—at least not based on the data currently available. Also it might serve as a reminder that the nonlinearity I employ in my psychophysical model does not directly map to the nonlinearity of neurons, although they use the same basic form.

If the connection to neuronal processing was closer, I could use the typically employed noise forms from physiology. In physiology noise is typically modelled as Poisson noise or variations of it with different factors between mean response and variance, or with additional variance shared between units (Goris, Movshon, & Simoncelli, 2014). For my model however, it is unclear how many neurons a channel response at a single pixel represents, and on which level of the model the noise relevant for a task is induced. Thus I believe that modelling the noise as Gaussian is warranted for simplicity.

I include no noise correlations in my model—it was simply unnecessary to add this additional "complication" in order to fit our psychophysical data. Including noise correlations in the model is computationally far from trivial, caused by the sheer number of activities which could be correlated. Furthermore, having to decide which channel responses should be correlated would add many additional degrees of freedom not constrained by psychophysical data. This does not argue against noise correlations, of course, but only that adding more uncorrelated noise adequately mimics the effects of these correlations for my purposes.

Processing heterogeneity

Like all previous spatial vision models—image-computable or not—I did not model the diversity of V1 neurons (and, presumably, psychophysical channels). For computational efficiency all the channels in my model have the same bandwidths, i.e. all neurons have the same receptive field, scaled and rotated to adjust their preferred spatial frequency and orientation. In contrast, V1 neurons have diverse bandwidths (De Valois, Albrecht, & Thorell, 1982; Ringach et al., 2002), which seems to be adaptive for natural scenes (Goris et al., 2015). Also all channels in my model cover the image with constant and equal density, although in truth V1 neurons seem to be sparser and the number of neurons differs between different spatial frequencies and orientations, which manifests itself in the psychophysical oblique effects (Furmanski & Engel, 2000; B. Li et al., 2003). Changing the density of neurons might be adaptive to concentrate resources on frequent stimuli and to implicitly represent the prior distribution over stimuli (Laughlin, 1983). However, as for the simple Gaussian noise approximation discussed above, my simplified model appears sufficient to capture human behaviour in response to classical psychophysically employed stimuli.

2.4.3 *Limitations of the presented model*

Although I tried to closely represent the concepts realised in classical spatial vision models, I did not include all ideas in my model for computational simplicity. Phrased negatively, I excluded substantial areas of spatial vision, as I discuss below.

Temporal dynamics, colour and stereo

I restrict my model to static, grey-scale luminance images projected onto a single cyclopean fovea for the reasons I gave in the introduction. This excludes any kind of temporal changes beyond the very coarse separation by presentation duration I made in my model. A true processing of stimuli over time would go beyond our current computational capabilities. Nonetheless it is worth highlighting that temporal processing was investigated and seems to be explainable by two or maximally three temporal channels (Watson, 1986; Watson & Nachmias, 1977). However, I am not aware of a combination of these models for temporal processing with masking or discrimination models. Furthermore, luminance images exclude colour processing, which requires considerably more complex models of the optics to include chromatic aberrations (Bedford & Wyszecki, 1957; Charman & Jennings, 1976) and of the retinal sampling, adaptation and processing, which differ between colour channels (Brainard, 2015). Additionally cortical processing of colour is understood less completely (Gegenfurtner, 2003). Finally luminance images contain no depth information, which relieves us from explicitly modelling 3D scenes, the optical effects on objects outside the focal plane and binocular vision. Modelling binocular vision is possible, but results in considerably more complex psychophysical models (Baker, Meese, & Georgeson, 2007; Georgeson et al., 2016; Legge, 1984a, 1984b; Meese et al., 2006). The additional complexity arises, because human observers do not only non-trivially combine the binocular input into one combined image, but can also perceive disparity (spatial shift between eyes) and lustre (contrast differences between eyes). Under dichoptic presentation these additional channels can lead to interesting unintuitive results (e.g. May & Zhaoping, 2016).

Adaptation

Our model includes no adaptation effects yet. This means that some classical psychophysical datasets are not within the scope of my model (Blakemore & Campbell, 1969, for example). Some abstract models (Foley & Chen, 1997; Goris et al., 2013; Meese & Holmes, 2002, for example) contain adaptation and discuss which parts of the model adapt to what kind of stimuli. However, adaptation would at least require additional input besides the stimuli to be discriminated and depends considerably on the duration of the adaptation stimulus and the interval between adaptation and test stimuli. Thus adaptation in an image based model would require substantial additional work, and would perhaps best be tackled after a model with adequate temporal dynamics exists.

Peripheral vision

I restrict myself to a purely foveal model, and thus to a model with uniform processing and sensitivity. Peripheral vision differs from foveal vision already in the optical quality (Jennings & Charman, 1981; Navarro, Williams, & Artal, 1993; Williams, Artal, Navarro, McMahon, & Brainard, 1996) and retinal processing—at least by the sampling density (Curcio & Allen, 1990; Curcio, Sloan, Packer, Hendrickson, & Kalina, 1987). Additionally, the interactions between channels, which I model in the normalization step, are different

in the periphery (Xing & Heeger, 2000). More generally, higher level restrictions like crowding (Whitney & Levi, 2011) play a larger role in the periphery, presumably due the stronger information reduction and the faster growth in peripheral receptive field size (Gattass, Sousa, & Gross, 1988; Rosenholtz, 2016). Hence a detailed modelling of the periphery would require a considerable effort beyond my current model.

Additional Tasks

In this Chapter I evaluate my model exclusively on discrimination data, and I cover a broad range of psychophysical data, but, of course, not all of it. Obvious omissions are data from direct estimation tasks ("What was the orientation of the grating?") as well as classification tasks ("Was the grating tilted left or right?"), because my model cannot deal with data from such tasks in its present form. Clearly, such tasks are important and have been used to investigate models of early visual processing (Meng & Qian, 2005; Solomon, Felisberti, & Morgan, 2004). Such tasks could be implemented as a different type of decoding based on the model representation. To explain biases in human perception explanations of these effects might require the inclusion of prior believes about the categories (Girshick, Landy, & Simoncelli, 2011) or deviations from optimal decoding.

2.4.4 *Applications in and beyond spatial vision*

On the one hand I hope to facilitate investigations into the details of spatial visual processing, using my model as a starting point or basis. Further developments are still necessary, not least to address the limitations and controversial design choice I discuss above. To further investigate spatial vision, image-computable models can be applied to a much wider range of existing data and allow the generation of optimized stimuli to differentiate different models, as I demonstrated in section 2.3.6. In addition, image-computable models allow direct comparisons to normative theories, as I have started on a small scale in this chapter. Whatever normative ideas might arise in the future, it can be assessed whether my spatial vision model optimises the proposed measures.

On the other hand, going beyond early spatial visual processing, my model might help with the development of mechanistic models of mid- or high-level visual processing by providing a psychophysically sound basis in which to represent images beyond pixels. Using a sound early processing model, I conjecture, might improve the match between mid- and high-level vision models and human perception. One clear target for such endeavours are convolutional DNNs in object recognition (Kriegeskorte, 2015; LeCun, Bengio, & Hinton, 2015; D. L. Yamins & DiCarlo, 2016).

Finally a working spatial vision model might have practical applications as an image quality metric as was the original intention of Teo and Heeger (1994). Later image quality metrics like the structural similarity metric (SSIM, Wang et al., 2004, 2003) claim to go beyond error visibility, but arguably getting error visibility right would be a good start as well. As it is currently demonstrated for the Normalized Laplacian Pyramid (Laparra et al., 2016), such image quality metrics can then be used to optimize the display of images to make it match the perception of the original (Laparra et al., 2017).

3

LIKELIHOOD-BASED EVALUATION OF DYNAMICAL COGNITIVE MODELS

Perhaps the principal value of this approach will be to facilitate understanding and use of likelihood functions as such, in the light of the likelihood principle, by relating them to concepts and techniques more familiar to many statisticians.

Birnbaum (1962)

In this chapter I present the likelihood-based evaluation methods I developed for general dynamical models. This model class especially encompasses most eye movement models allowing arbitrary causal dependencies between fixations, as long as the model's prediction for the next fixation can be calculated based on the previous eye movements. The measures of model performance will be essential for the evaluation of models in Chapter 4. The content of this chapter was published in highly similar form in the following article:

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, 124(4), 505–524.

3.1 INTRODUCTION

The broad class of dynamical cognitive models (Van Gelder, 1998) provides a powerful framework for explaining behavioral data. This modelling approach has been particularly successful in sensorimotor control. For example, an early paradigmatic model was proposed by Haken, Kelso, and Bunz (1985) who introduced coupled non-linear oscillators as a mathematical model for phase transitions in human finger movements. Another general theory was proposed by Erlhagen and Schöner (2002) who introduced a flexible framework of movement preparation based on dynamical equations for the temporal evolution of neural fields that specify motor actions in space and time. With their decision field theory, Busemeyer and Townsend (1993) developed a dynamical framework for decision making in uncertain environments. These representative examples indicate the broad range of dynamical models in cognitive science.

A strength of the dynamical approach is to generate specific predictions including the dependencies between different data-points over time. This however implies that the statistical treatment of dynamical models requires the comparison of model predictions for time-ordered and interdependent data, which complicates parameter identification and model comparison. As a result, dynamical models are often handled with heuristic and approximate methods. Here I discuss an alternative to these heuristic approaches, namely a statistically well-founded analysis based on the likelihood framework.

An important application of the dynamical framework is the modeling of eye movements. Human observers move their eyes three to four times per second to shift gaze to regions of interest within a given visual scene (Henderson, 2003; Yarbus, 1967). Eye movements are important, since high-acuity vision is limited to the fovea, a small region with a spatial extension of about 2 degrees of visual angle (Nicholls et al., 2012; von Helmholtz, 1924). The analysis of fixated regions permits conclusions on the type of features that attract our gaze. For eye movements in natural scenes, *saliency models* concentrate on predicting the fixation density for large datasets (Itti & Koch, 2001). The density of fixations provides only information where people look regardless of serial order and durations of fixations. This research strategy turned out to be very successful and a range of saliency models was developed to predict fixation density for a given input image (Borji & Itti, 2013; Kienzle et al., 2009; Kümmerer et al., 2015).

Recently, there is an increasing interest in cognitive models that produce sequences of fixations, i.e., a *scanpath*, on a natural scene (Borji, Sihite, & Itti, 2014; Engbert et al., 2015; Le Meur & Liu, 2015; Zelinsky, 2008). Related models aim at a more complete explanation of the cognitive principles underlying the control of attention and eye movements during exploration of natural scenes. Statistical measures include simple statistics like the distribution of saccade lengths and angles between subsequent saccades (R. M. Klein & MacInnes, 1999; Smith & Henderson, 2009), but also more complex spatial statistics that

relate image properties to fixation density (Barthelmé, Trukenbrod, Engbert, & Wichmann, 2013) or to spatial correlation functions (Engbert et al., 2015).

In the traditional approach for the evaluation of scanpath models, researchers typically simulate scanpaths from their models and compare simulated data to experimentally observed scanpaths using a broad range of statistics (Le Meur & Baccino, 2013). The most common statistics are those associated with the observed experimental data (e.g., distributions of saccade angle and saccade amplitudes). Alternative methods are based on comparisons of scanpaths that include string comparison methods based on the Levenshtein distance (Levenshtein, 1966; von der Malsburg & Vasishth, 2011, for reading) or vector-based methods (Jarodzka, Holmqvist, & Nyström, 2010). However, each effect and each discriminating statistic for scanpaths evaluates different aspects of the models. Thus, ranking of model performance depends critically on which effects are investigated and which statistics are applied. None of the statistics used so far quantifies the general agreement between models and experimental data in a dynamical framework.

For saccade generation in dynamical cognitive models, a spatiotemporal map of activations (Erlhagen & Schöner, 2002) is built-up according to dynamical evolution equations (e.g., Jackson, 1992). When a saccade target is needed, the activation map is read out to generate a target with a probability that equals the relative activation as determined by the map at the time of saccadic selection. I study a dynamical model of scanpath generation for eye movements in scene viewing (Engbert et al., 2015). While I focus on this concrete example to illustrate the procedures of model parameter identification and model comparison, the model only serves as a representative example for the broad class of dynamical cognitive models that are developed for the prediction of sequences of discrete motor actions.

In this chapter, I investigate the application of the *likelihood function* as a statistical measure of model performance. The likelihood function of a model M is the probability that a given set of experimental data was generated by the model and a corresponding set of model parameters $\boldsymbol{\theta}$. Therefore, the likelihood function for a given model depends on the data set and the set of model parameter values that specify the model's behavior. The likelihood is the most widely used measure of model performance in mathematical statistics (Bickel & Doksum, 1977; Cox, 2006). However, because its numerical computation is believed to be difficult, the likelihood is not yet part of the standard toolbox for dynamical models of cognition. Solving likelihood computation for dynamical models of cognition is potentially very important, since likelihood is the starting point for many additional concepts of statistical inference about model parameters and comparisons between different models, including Bayesian inference (Jaynes, 2003).

The likelihood can be computed whenever the model can generate the observed data with a certain probability that is non-zero. This is already guaranteed, if the probability for the next datum can be calculated given the previous data and is greater than zero for any observed datum. This means that the likelihood approach can be applied to an extremely broad class of models.

To investigate how the analysis of dynamical models can benefit from the likelihood approach, I demonstrate numerical computations for the recently published *SceneWalk* model of scanpath generation in natural scene viewing (Engbert et al., 2015). The general motivation for modelling human scanpaths is to derive the rules for the sequential deployment of overt attention (i.e., gaze position) in a natural scene-viewing task. The *SceneWalk* model starts from a given spatial distribution of fixation positions (an *empirical saliency map*). Thus, I assume to have perfect knowledge about saliency (up to differences between observers). This is not a strong limitation, since the model could eas-

ily be combined with one of the successful saliency models (see Borji & Itti, 2013, for an overview). Thus, our modelling goal is to reproduce the key statistics of human scanpaths (e.g., distribution of saccade lengths and spatial correlations) for a given image, when the time-independent 2D distribution of fixation positions is known to a good approximation.

3.2 LIKELIHOOD COMPUTATION FOR DYNAMICAL MODELS

Definition of likelihood function

The fundamental theoretical concept for our approach is the likelihood $L_M(\boldsymbol{\theta}|\,\text{data})$ of a model M with parameters $\boldsymbol{\theta}$ given a specific set of experimental data, which is defined as the conditional probability density f_M for observing the data in the context of model M specified by parameters $\boldsymbol{\theta}$, i.e.,

$$L_M(\boldsymbol{\theta}|\,\text{data}) = f_M(\text{data}\,|\boldsymbol{\theta}) \approx \frac{P_M(\text{data}\,|\boldsymbol{\theta})}{(\Delta A)^N}\,. \tag{9}$$

In our case, data are given by a sequence of fixations, for which our models shall predict a density one after another. Each of these densities can be approximated by the probabilities to observe the fixations exactly on a discrete grid, divided by the area each gridpoint represents resulting in a denominator of $(\Delta A)^N$ for N fixations. I will stay with this grid approximation to all likelihoods in this article, as many models are themselves defined on grids, including saliency models and the SceneWalk model that I investigate in this chapter. The grid approximation simplifies numerical computations, since this probability is always defined and all integrals reduce to summations over grid points.

Furthermore I set $\Delta A = 1$, measuring area in grid points, which works, because all models that I aim to compare to each other make predictions on the *same* grid of possible fixation locations. Measuring the area in grid independent units (cm, pixels, degrees of visual angle, etc.) in principle enables comparisons between models, which are defined on different grids. Using a coarser grid implicitly blurs model predictions for eye movement models and a blurring of the final predictions may change performance considerably (Judd et al., 2009). Thus I think it is preferable to convert all model predictions to the same grid making all necessary conversions explicit.

The likelihood quantifies how well a model describes the data and is the most common criterion for model evaluation in mathematical statistics. Therefore maximizing the likelihood of a given dataset by optimizing model parameters[1] is a straightforward approach to model fitting. Applicability of the likelihood approach depends on both the structure and complexity of a model M, i.e., whether the likelihood can be computed exactly (analytically or via numerical simulation of the model) or whether one needs to introduce further approximations. If it is not practical to compute the likelihood, likelihood-free strategies for parameter estimation and model comparison have been proposed as an alternative (see Discussion).

1 I only consider finite dimensional parameters and models in this chapter. I know of no non-parametric models for scanpath generation. A non-parametric model increases the complexity of the analysis considerably. If the reader is interested in this there is a broad literature on non-parametric statistics in both Frequentist (Conover, 1980) and Bayesian statistics (Gershman & Blei, 2012)

3.2.1 *The likelihood for dynamical models based on discrete observations*

To calculate the likelihood for dynamical models based on time-ordered experimental data and, specifically, for the SceneWalk model of eye movements in scene viewing (Engbert et al., 2015), we split the likelihood into a product of probabilities for all fixations $f_i = (x_{f_i}, y_{f_i})$ given the previous fixations $f_1 \dots f_{i-1}$ in the sequence, i.e.,

$$\begin{aligned} L_M(\boldsymbol{\theta}|\,\text{data}) &= L_M(\boldsymbol{\theta}|f_1,\, f_2,\, \dots,\, f_n) \\ &= P_M(f_1) \prod_{i=2}^{n} P_M(f_i|f_1,\, \dots,\, f_{i-1}, \boldsymbol{\theta})\ , \end{aligned} \tag{10}$$

where $P_M(f_1)$ is the probability of the initial fixation starting at time $t = 0$, which can be given by the experimental design or the model. The conditional probabilities $P_M(f_i|f_1 \dots f_{i-1}, \boldsymbol{\theta})$ can be computed by enforcing the model to generate the sequence of fixations $f_1, \dots, f_{i-1}$ to obtain the probability for the i^{th} fixation f_i. This is possible in dynamical models which generate a continuous-time activation map u that translates into a fixation probability π to place the next fixation at position f_i at time t. Thus, we can read out the probability for the next fixation from the map u, Eq. (18), via the transformation given in Eq. (21). During numerical simulation, we force the model to generate a particular scanpath prescribed by the data $f_1,\, f_2, \dots$, which translates into a certain probability at each iteration and reduces the necessary computations to a single model run for a given scanpath. This procedure is illustrated for the first fixations on an image in Figure 18.

For practical purposes, it is advantageous to use the logarithm of the likelihood (log-likelihood):

$$l_M(\boldsymbol{\theta}|\text{data})) = \log(L_M(\boldsymbol{\theta}|\text{data})) \tag{11}$$

$$= \sum_{i=1}^{N} \log(P_M(f_i|f_1 \dots f_{i-1}, \boldsymbol{\theta})) \tag{12}$$

The log-likelihood can be calculated and optimized more easily, since it transforms the products over observations into sums of terms and scales numerical values to a more feasible range.

The log-likelihood characterizes model performance on the whole dataset, in the current case the fixation sequence or scanpath. Therefore, the log-likelihood of a scanpath given a model depends on the length of the sequence or number of fixations. To obtain a number that is easier to compare between different realizations of scanpaths, it is more informative to compute the log-likelihood per fixation, which turns out to represent a sensitive measure of model performance as the log-likelihood is added up over all fixations in a given sequence.

Thus, effectively, I compute the average probability of an observed fixation, calculating the average as a geometric mean. However, I express all likelihoods on a logarithmic scale. When the $\log_2$ is used as I do in this chapter, the unit of the log-likelihoods is a *bit*. A difference of 1 bit between two log-likelihood values thus indicates that the corresponding likelihoods differ by a factor of two.

A log-likelihood of zero indicates that the model predicted the observed data exactly and with probability one. This is a limiting case and certainly not a realistic scenario for typical cognitive models. Almost always models predict a distribution over multiple possible outcomes, which each have smaller probabilities than one. Therefore, log-likelihoods are

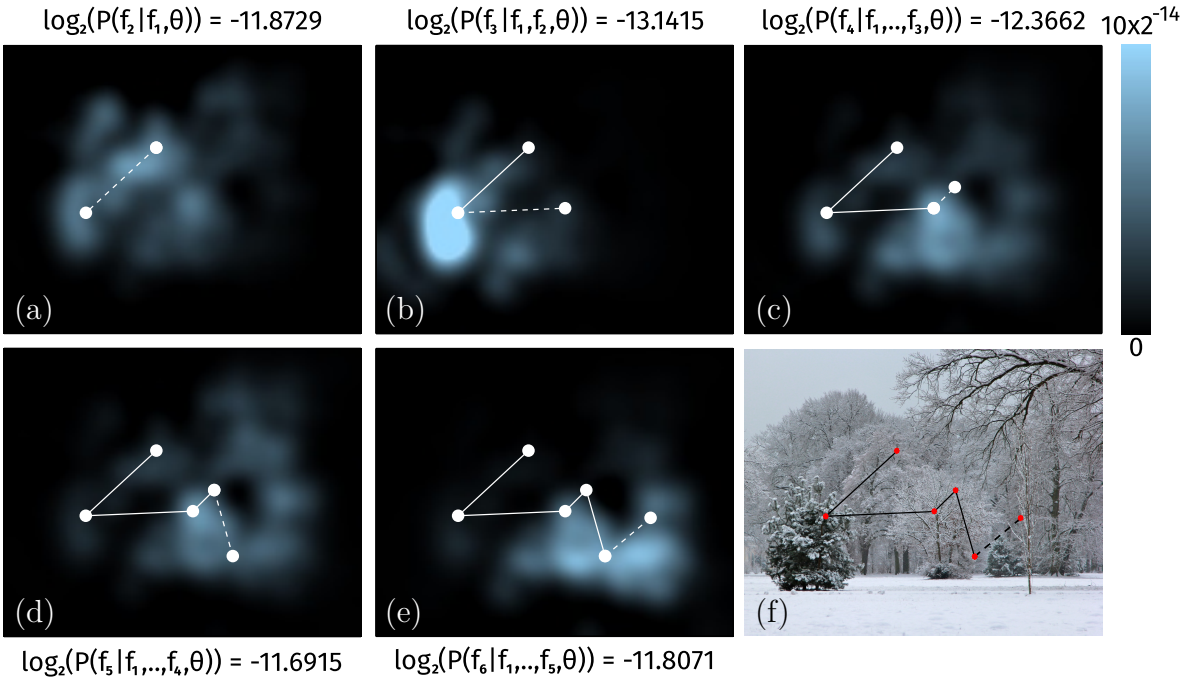

Figure 18: Numerical calculation of the likelihood for an example of a fixation sequence. **(a)-(e)** Visualization of the probabilities of the first 5 fixations from a sequence as predicted from the SceneWalk model. We compute the probability $P(f_i|f_1 \dots f_{i-1}, \boldsymbol{\theta})$ of the next fixation, which the human observer actually generated and force the model to choose the fixation location accordingly. With this new location we can calculate the probability distribution for the next saccade and can thus iterate through the observed scanpaths and calculate their probabilities given by the model and its parameter values. **(f)** The presented image with the scanpath overlayed.

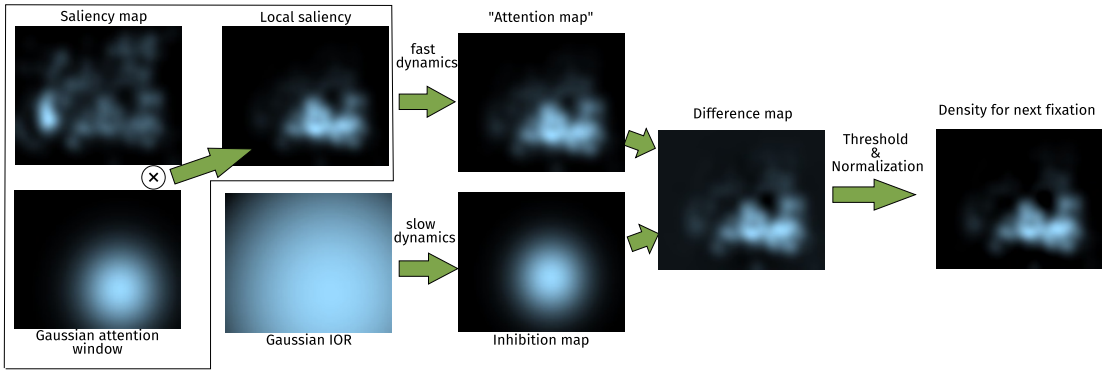

Figure 19: Schematic illustration of the SceneWalk model (Engbert et al., 2015). The temporal evolution of two independent processing streams for attention and inhibition-of-return is combined into the time-dependent potential $u(x,t)$ that determines the next saccade target. The saliency map is weighted by a Gaussian (attentional window) placed at the current fixation. The resulting local saliency map is used as the input for the build-up of activation in the attention map. An inhibition map is subtracted, which builds up more slowly using a constant-shape Gaussian around the current fixation as input. Finally, thresholding and normalization yield the final distribution $u(x,t)$ for the probabilistic selection of the next saccade target.

almost always negative. Indeed, the log-likelihoods I calculate below will usually be in the range between $-10\frac{\text{bit}}{\text{fix}}$ and $-20\frac{\text{bit}}{\text{fix}}$.[2]

3.2.2 *Model details*

For the analysis of the likelihood of the SceneWalk model we need to compute the probability for the next fixation, given all previous fixations in a given trial. In this section I describe how the SceneWalk model computes this probability distributions. Explaining these calculations requires a short recap of the model internals and I will take the opportunity to describe the details of some variants of the model I will use to exemplify model comparisons below.

The SceneWalk model is based on two independent processing streams for excitatory and inhibitory aspects of saccade planning that are related to attentional deployment (Itti & Koch, 2001; Itti et al., 1998) and inhibition-of-return (R. M. Klein, 2000; R. M. Klein & MacInnes, 1999), respectively (Fig. 19). The excitatory pathway starts with a given fixation density (empirical saliency), which is multiplied with a Gaussian attention window around the current fixation location resulting in a local saliency map. This localization step serves as a first-order approximation to the peripheral loss in available information, cortical processing, and visual attention. For the inhibitory pathway we start with a simple Gaussian around the current fixation marking the currently visited area. The local saliency and the inhibitory Gaussian are both implicitly time-dependent through changes of gaze position.

For a current fixation position $\mathbf{x}_f = (x_f, y_f)$ we first compute the two Gaussian distributions centred at $\mathbf{x}_f$ on a grid of size $L \times L$. The attentional pathway uses a Gaussian

2 Note that these reference values are specific for our choice of grid and area unit, such that they cannot be compared to values obtained with a different grid or area unit. Especially, densities and thus likelihoods can be larger than 1 and log-likelihoods larger than 0, depending on the measure of area chosen.

aperture G_A with standard deviation σ_A to access the static empirical saliency map. The pathway for inhibitory tagging uses a Gaussian G_F with standard deviation σ_F to build-up inhibition that drives the model to new regions of the visual field. For a grid position (x, y) these Gaussians are given by

$$G_{A/F}(x, y; x_f, y_f) = \frac{1}{2\pi\sigma_{A/F}^2} \exp\left(-\frac{(x - x_f)^2 + (y - y_f)^2}{2\sigma_{A/F}^2}\right) . \tag{13}$$

Next, we define the change over time of the attention map $A(t) = \{A_{ij}(t)\}$ and the fixation map $F(t) = \{F_{ij}(t)\}$ with indices $1 \leq \{i, j\} \leq L$ running over the whole image. Two parameters ω_A and ω_f scale the rates of activation change in the two maps and we require the given time-independent salience map $S = \{S_{ij}\}$ and the Gaussians G_A and G_F from equation (13):

$$\frac{\mathrm{d}A_{ij}(t)}{\mathrm{d}t} = -\omega_A A_{ij}(t) + \omega_A \frac{S_{ij} \cdot G_A(x_i, y_j; x_f, y_f)}{\sum_{kl} S_{kl} \cdot G_A(x_k, y_l; x_f, y_f)} \tag{14}$$

$$\frac{\mathrm{d}F_{ij}(t)}{\mathrm{d}t} = -\omega_F F_{ij}(t) + \omega_F \frac{G_F(x_i, y_j; x_f, y_f)}{\sum_{kl} G_F(x_k, y_l; x_f, y_f)} , \tag{15}$$

where the $\sum_{kl}$ symbol denotes the sum over all grid-points (k, l).

These evolution equations were formulated as difference equations in Engbert et al. (2015). However, I moved to differential equations here, as they can be solved analytically. By solving Eqs. (14 & 15), I can exploit the fact that the input $G_{A/F}$ changes only due to saccadic gaze shifts $\mathbf{x}_f \mapsto \mathbf{x}'_f$. The solution of the differential equations for initial maps A_0 and F_0 at the start of the fixation at time t_0 are given as

$$A(t) = \frac{G_A S}{\sum G_A S} + e^{-\omega_A (t - t_0)} \left(A_0 - \frac{G_A S}{\sum G_A S}\right) \tag{16}$$

and

$$F(t) = \frac{G_F}{\sum G_F} + e^{-\omega_F (t - t_0)} \left(F_0 - \frac{G_F}{\sum G_F}\right) , \tag{17}$$

where indices have been dropped to simplify the representation. As a consequence of the linear dynamics of the maps, the solutions describe exponential change from the map represented at the beginning of the fixation towards the input map. Using these equations one can calculate the activities at the end of the fixation directly. Another advantage is that this formulation prevents temporal discretization errors (in the original model, a 10 ms temporal discretization was used, see Engbert et al., 2015, for details).

At the first fixation the maps in the model need to be initialized. The original model was initialized with zero activities of the maps for attention and inhibitory tagging. For short durations of the first fixation, however, this led to unintended behavior, as the maps are normalized. Small activations on the maps are amplified by the normalization which introduces unwanted starting effects. To prevent this problem of the model's initial conditions, I prepared the maps with a uniform distribution of sum one and adjusted the magnitude of the input such that the equilibrium size of the maps was normalized to one as well. Thus, the sum of activation of the attention map and of the map for inhibitory tagging remains at a constant value of one throughout each simulated trial.

Finally, the two independent activation maps $A(x, t)$ and $F(x, t)$ are combined into a map $u(x, t)$, which is defined as the difference of the attention and inhibition maps after thresholding and normalization. To obtain a flexible relative weighting within each map, numerical values of activations are raised to power λ for the attention map A and to power γ for the fixation map F, respectively. Next, each map is normalized to unit sum

(Carandini & Heeger, 2012). Finally, the map for inhibitory tagging is multiplied by a factor c_F and subtracted from the attention map. As a result, we obtain a time-dependent potential $u_{ij}(t)$ for target selection:

$$u_{ij}(t) = \frac{[A_{ij}(t)]^\lambda}{\sum_{kl}[A_{kl}(t)]^\lambda} - c_F \frac{[F_{ij}(t)]^\gamma}{\sum_{kl}[F_{kl}(t)]^\gamma} \,. \tag{18}$$

Note that I introduced the factor c_F as an additional parameter, which was not present in the original model (Engbert et al., 2015).

Taking a power of the map at each point changes not only the weighting between different peaks, but also shrinks or widens the individual peaks. Therefore, to obtain parameters which represent the size of the final influence and are thus easier to interpret, I re-parametrized the model using the following equations:

$$\lambda\sigma_A'^2 = \sigma_A^2 \qquad \gamma\sigma_F'^2 = \sigma_F^2 \tag{19}$$

Thus σ_A' and σ_F' are the standard deviations the Gaussians would have if they were mapped through the nonlinearity directly.

Normalization. To obtain a probability distribution from $u_{ij}(t)$, the potential is normalized to be positive and to have a unit integral over the whole image. Compared to the published version of the model (Engbert et al., 2015), I changed several aspects on the normalization of u and on the initialization of the maps at the beginning of a trial, which are explained in the following. In the normalization procedure of the original model, negative values of the potential $u_{ij}(t)$ implied probability zero to select position (i, j) as the next saccade target. However, this is an unrealistic assumption in the model, since experimental data do not indicate regions which are never selected as a saccade target. I changed the model accordingly. First, we define a function which continuously maps u to an intermediate u^*, which is positive everywhere, i.e.,

$$u^*(u) = \begin{cases} u & u > 0 \\ 0 & u \leq 0 \end{cases} \tag{20}$$

In a second step we compute a mixture with a uniform distribution using a weighting factor ζ to obtain the probability $\pi(i, j)$ for each position on the lattice to be selected as the next fixation target,

$$\pi(i,j) = (1-\zeta)\frac{u^*_{ij}}{\sum_{kl} u^*_{kl}} + \zeta\frac{1}{\sum_{kl} 1} \,. \tag{21}$$

This formulation maps the original function u to a probability on the map, which always returns a positive probability ($\geq \zeta / \sum_{kl} 1$) for any next fixation. Furthermore, areas with high u are not further distorted by this mapping, such that relative weightings from the original empirical saliency map are kept.

The distribution $\pi(i, j)$ directly represents the probability of a specific grid-point to be the next fixation target, given the previous fixations, i.e., the map to be used in the likelihood calculation described in Equation 10 and illustrated in Figure 18 completing our description of the likelihood calculation for the SceneWalk model.

3.2.3 *Competing models*

Below I will compare the SceneWalk model to some other models, whose details are described in this section.

Non-dynamic benchmarks. First, I compare the performance of the SceneWalk model to non-dynamical models that represent limiting cases for saliency evaluation: An image independent spatial bias and empirical saliency. The image independent spatial bias mostly represents the central fixation bias (Buswell, 1935; Tatler, 2007)—the experimental observation that observers initially direct their gaze positions toward the image center. A corresponding model can be realized as an image-independent kernel density estimate of all fixations of the full set of images. The empirical saliency model represents the optimal prediction of fixation positions from other observers generated as a kernel density estimate as well, using fixations on the tested image only. Additionally, I implemented a model which generates a uniform distribution over the full image as a null model setting an absolute zero point on our log-likelihood scale.

A model without inhibition. As a first dynamical model to compare to, I chose a model without inhibition, to test whether this part of the model is necessary as the influence of inhibition of return on scene viewing behaviour has been challenged recently (Smith & Henderson, 2009). To implement this model I simply set $c_F = 0$ in our original model removing the influence of the inhibitory pathway. As u then cannot become negative anymore, I also replaced the mapping from u to u^* with the identity. As a consequence, all parameters of the inhibitory pathway are superfluous in this model, such that we are left with only 4 parameters for this model: $\omega_A, \sigma_A, \lambda$ and ζ.

Divisive inhibition model. The original SceneWalk model implements a subtractive inhibition. However, there are no strong reasons, why this inhibition should be subtractive. An alternative and common model of interaction is divisive inhibition (Carandini & Heeger, 2012). To test this alternative form of combining the two maps, I changed the formula for u to:

$$u_{ij}(t) = \frac{[A_{ij}(t)]^\lambda}{c_F^\gamma + [F_{ij}(t)]^\gamma} \tag{22}$$

As for the model without inhibition, the variable u cannot become negative. Again, I replaced the mapping from u to u^* with the identity. This way to combine excitation and inhibition has the same number of parameters as the original subtractive formulas. Thus we are left with 8 parameters as for the original model.

3.3 ESTIMATION OF MODEL PARAMETERS

As it is common practice the previous approach to the estimation of model parameters of the SceneWalk model was based on minimization of an ad hoc loss function that included gaze positions and saccade lengths as measures of model performance (see Appendix in Engbert et al., 2015). First, the squared differences between densities of gaze positions from experimental and simulated data were computed using 2D bins for discretization. Second, the experimentally observed and simulated saccade lengths were compared via squared differences from bins of the distributions. The sum of both measures was minimized to obtain parameter estimates.

However, there were several problems associated with this approach that motivated me to develop an alternative framework. First, our earlier approach worked for a limited set of parameters only. Some of the parameters had to be fixed at plausible values. These fixed parameters included important parameters, for example, normalization exponents of the dynamic activation maps, which are critical for the spatial correlation functions the SceneWalk model was intended to reproduce. Second, the qualitative model analyses necessary to find useful and plausible values for the fixed parameters required time-consuming

hand-selected model runs. Third, the earlier fitting approach based on a subset of hand-selected fixed parameters and estimates from minimization of an ad-hoc loss-function could not guarantee reliable or consistent estimates and was missing a statistical justification. Moreover, confidence intervals of the model parameters were inaccessible and were, therefore, replaced by an ad-hoc indicator of errors of parameter estimates derived from multiple runs of the minimization algorithm. Due to these shortcomings of the earlier approach, I set out to develop an improved strategy for parameter estimation that would be statistically well-founded, reliable, and efficient in terms of computer time, while working for all parameters.

3.3.1 *Maximum likelihood estimation*

A tutorial on the MLE concept for model fitting is given by Myung (2003) in the context of mathematical models in psychology (see Hays, 1994, for a more general context). The general idea is to find the particular (vector-valued) parameter $\boldsymbol{\theta}$ that corresponds to the maximum of the likelihood function given the observed data. This parameter value is used as a parameter estimate and, therefore, termed *maximum likelihood estimate* (MLE).

Fitting models to data based on the likelihood has considerable statistical advantages over using other statistics for fitting (Myung, 2003). First, the likelihood guarantees sufficiency, i.e., raw data do not constrain the parameters more than the maximum likelihood criterion. Second, for the likelihood, there is asymptotic consistency, such that for large samples the estimate converges to the correct parameter value if the data were generated from the model. Third, the likelihood has asymptotic maximum efficiency, i.e., for large samples, there is no consistent estimate with smaller variance. Finally, the likelihood estimate is not changed by the re-parametrization of the model, which is known as parametrization invariance.

In numerical simulation models like the SceneWalk model, the maximum of the likelihood can be found using an optimization algorithm that evaluates the likelihood $L_M(\boldsymbol{\theta}|\text{data})$ varying the model parameters $\boldsymbol{\theta}$. Most optimization algorithms try to change the parameters gradually to improve the likelihood and can thus be trapped in local extrema, where the likelihood is higher than for surrounding parameter values, but not the globally best parameter value. If the global optimum is found, it must not depend on the specific optimization algorithm or starting position. Consequently it is common practice to run multiple optimizations with different starting positions. If one of the local extrema is clearly better than the others and the optimizations end up in clusters, one can be reasonably sure that one found the global optimum.

Alternatively the field of global optimization designs algorithms to find global minima. Two well known families of algorithms for global optimization are: Simulated annealing, which—inspired by the cooling of physical materials—first explores broadly and later allows less and less bad objective values settling near the optimum (Kirkpatrick, 1984; Kirkpatrick, Gelatt, & Vecchi, 1983), and the Genetic algorithm, which simulates a population of parameter values over generations in which points with high objective function values have higher probability to reproduce in the next generation (Goldberg, 1989; Holland, 1975; Houck, Joines, & Kay, 1995). Variants of both these algorithms are available for most higher programming languages like MATLAB or python. As a promising idea for the future the relatively recent meta-modelling approach aims to model our knowledge about the function gained so far and to conclude which points to sample to gain the most information about the optimum (Hennig & Schuler, 2012; Jones, Schonlau, & Welch, 1998; Villemonteix, Vazquez, & Walter, 2009).

For optimization of the parameters of the SceneWalk model I employed the genetic algorithm for global optimization as implemented in MATLAB (R2016a). I used 200 individuals on the logarithm of the parameters with a range from -10 to 10 corresponding to a range from 0.000 045 to 22 026 for the parameters. Subsequently I further optimized using the Nelder-Mead Simplex Algorithm as implemented as fminsearch in MATLAB. Using the standard settings for all other options these algorithms found the global maximum reliably, as confirmed by some standard optimization runs from random start positions, the sampling I did for Bayesian inference and the fits I computed for cross validation as described below.

3.3.2 *Bayesian inference*

If the likelihood $L_M(\boldsymbol{\theta}|\text{data})$ of the data can be computed for a given model M, then Bayesian inference (Gelman, Carlin, Stern, & Rubin, 2014; Marin & Robert, 2007, for overviews) is a viable method for parameter estimation. The main advantage of Bayesian inference in the current context is that it provides not only the best fitting parameter values, but also a full distribution of possible parameter values. Thus, there is information on which other parameter values could also explain the data and thus how well the parameters of the assumed model are constrained by given data. In Bayesian inference, the goal is the computation of a posterior distribution $P(\boldsymbol{\theta}|\text{data})$ that indicates the most probable parameter values $\boldsymbol{\theta}$ under the assumption of model M and the given data. Based on the likelihood $L_M(\boldsymbol{\theta}|\text{data})$ and a prior distribution $P(\boldsymbol{\theta})$, which describes our knowledge or beliefs about the parameters prior to data collection, the posterior distribution is computed as

$$P(\boldsymbol{\theta}|\text{data}) = \frac{L(\boldsymbol{\theta}|\text{data})P(\boldsymbol{\theta})}{\int_\Omega P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}}\,, \tag{23}$$

where, computationally, the main problem is that quantities of interest are usually integrals over the posterior $P(\boldsymbol{\theta}|\text{data})$ like the expected value of the posterior, its variances or correlations. To compute these integrals it is often necessary to use Markov Chain Monte Carlo (MCMC) methods (Brooks, Gelman, Jones, & Meng, 2011; Robert & Casella, 2013). These methods produce—sometimes weighted—samples from the posterior using only local evaluations of the likelihood and prior. These samples can then be used to replace integrals by sample means. This especially avoids the direct calculation of the denominator $P(\text{data}) = \int_\Omega P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}$, which in turn can be computed from the samples if one is interested in this value.

The most controversial aspect of Bayesian statistics is the choice of prior. The main reason is that the prior may serve very different functions in different situations.

The first most literal interpretation of priors is that they shall represent all available believes prior to the experiment. If one manages to formulate all prior believes into the prior distribution, the posterior represents the believes one should have after the experiment to do proper reasoning (Jaynes, 2003, Chapter 1). If one had an estimate of the parameters from some other experiment, or had any other kind of information what the parameters or predictions of the model should be, the prior offers a possibility to include this knowledge. In the absence of prior information the general recommendation is to use relatively broad uninformative priors to avoid biasing the conclusions too much. If a bias is unavoidable, then the recommendation is modified to use a prior which favours the opposite of the suspected conclusion to achieve a conservative analysis showing how well the data should convince a sceptic (Gelman et al., 2014, Chapter 2.8, Jaynes, 2003, Chapter 11 & 12).

The notion of an uninformative prior can be formalized mathematically, which leads to Jeffreys' priors (Jeffreys, 1946). Another mathematically preferable kind of prior are conjugate priors, for which the posterior has the same form as the prior (Gelman et al., 2014, Chapter 2.4) such that posteriors can be parametrized and analytically analyzed. Neither Jeffreys' priors nor conjugate priors are particularly relevant for the complex models I study here, as they are rarely known or even computable for highly complex models.

A second more objective interpretation is that the priors shall represent the actual distribution of parameters as close as possible. In this interpretation, which is popular in machine learning, the prior becomes part of the model to be evaluated. The better the prior represents the distribution of parameters needed to fit data, the better it is. Obviously such evaluations require multiple instances for which a parameter is fitted. Once one starts to adjust the prior to fit some data this approach becomes essentially equivalent to hierarchical models which I discuss below.

Prior assumptions on parameters also represent a helpful tool to include information obtained from other experiments and other knowledge (e.g., physiological constraints) or to *regularize* the model, which is a general expression for preferring some parameter values of the model over others, if both parameter values explain the data equally well. The term regularization is used usually in Frequentist contexts and justified as a means to stabilize model fitting when the parameters are not sufficiently constrained by the data.

For regularization purposes one typically differentiates whether parameter values are only considered less likely or impossible. Only the former is usually called regularization, the later is usually called constrained estimation. This distinction is mainly necessary because once there are areas of parameter space which are impossible the algorithms for optimization or sampling need to be changed. For the effect of the priors on the model this is a more gradual distinction. While it is usually discouraged to entirely exclude parameter values a priori, i.e., to set their prior probability to 0, very small prior probabilities will have the same effect on the model predictions and parameter fits.

The different aims for priors partially work against each other. To regularize or to include prior knowledge helps mostly if the parameters could not be constrained well by the data at hand, i.e. when the prior excludes parameters which could fit the data convincingly as well. When doing this one can obviously not interpret the posterior as information how well these parameters are constrained by the data. Thus different aims might require different priors for the same model and data.

As I do not require regularization and have little to no prior information about the parameters of the model I investigate, I chose an extremely broad prior not to influence our parameter estimates. I assume a log-normal distribution with a standard deviation of 30 units (log-space) around 0 (in log-space).

3.3.3 *Results on model parameter estimation*

For the SceneWalk model, I used the same dataset as in the original article (Engbert et al., 2015). In the experimental data, gaze positions were recorded via eye tracking from 35 human observers in a memorization task. Experimental stimuli consisted of 15 natural images and 15 texture images, where the latter are photographs of relatively homogeneous textures like grass or a stone wall.

The numerical optimization of the model parameters required less computation time than the original fitting method, as the likelihood objective is not stochastic, although I

fitted four more parameters (the pooling exponents λ and γ, the weighting of the inhibitory map c_F and the weight of the uniform map in the mixture ζ).

The results of the Maximum Likelihood estimation are listed in Table 2. As they agree with values from Bayesian estimation I shall discuss their meaning after explaining the origin of the Bayesian estimates.

To perform Bayesian inference about the parameters of the SceneWalk model, I sampled the posterior distribution with a Metropolis Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). A hand-tuned multivariate Gaussian proposal distribution was chosen to have a covariance matrix roughly proportional to the covariance of the sampled distribution and to reach an acceptance rate of roughly 25% as recommended as optimal for Gaussians by Gelman, Roberts, and Gilks (1996). I restricted myself to reproduce the diagonal of the covariance Matrix, i.e., to the variances of the individual parameters, and 3 particularly strong covariances, between σ_A and σ_F, C_F and λ and C_F and ζ respectively. Using this scheme I sampled three chains with 50 000 samples each starting with a small displacement from the MAP estimate. I then discarded the first 1000 samples as burn in, which covered the initial transient back towards the MAP in all parameters.

First, I checked that my sampling algorithm converged using the $\hat{R}$ statistic (Brooks & Gelman, 1998; Gelman & Rubin, 1992), which quantifies how large the variance between chains is compared to the variance within the chains, i.e., whether the chains sampled different regions. The $\hat{R}$ statistic is always greater than one and, when the chains under analysis converged to the same stationary distribution, the $\hat{R}$ statistic should be close to one. For my chains I obtained values in the range from 1.00 to 1.06 for different parameters and a value of 1.06, when $\hat{R}$ was computed as a multivariate statistic. I thus concluded that our chains converged to their common stationary distribution, which I also confirmed by investigating visually and by comparison of the distributions obtained from the three independent chains.

Next I checked that my chains mixed sufficiently well, i.e., I tested that the samples were sufficiently uncorrelated with each other and, therefore, that the samples provide an adequate representation of the posterior distribution. The mixing property was analysed via the effective sample size, which is an estimate of the number of independent samples one would need to get an equally good representation of the posterior. This estimate is computed from the autocorrelation of the chain for each individual parameter. As a result, I obtained an estimate of the effective sample size for each parameter, although the true efficiency of the sampling algorithm is a single quality of the method. For our chains, the effective sample sizes turned out to range from 624 to 22806 for the different parameters. This indicates that our sampling algorithm provides at least the information of a few hundred samples, which I considered to be sufficient for our purposes.

However, our findings on the effective sample size also indicate that the Metropolis Hastings algorithm could probably be improved in efficiency as its sampling efficiency (effective sample size divided by the number of drawn samples) was less than 1%. When the algorithm is well tuned to the problem, a sampling efficiency of several percent can be reached (Gelman et al., 1996).

The sampled posterior distributions are displayed in Figure 20. The distributions clearly indicate the most likely values of the parameters. All parameters except for the decay of the excitatory map ω_A and the exponent γ were well constrained by the data. Their posterior marginals concentrate on a range of $\leq \pm 10\%$ around the best fitting values and are much narrower than the prior (± 10 log-units).

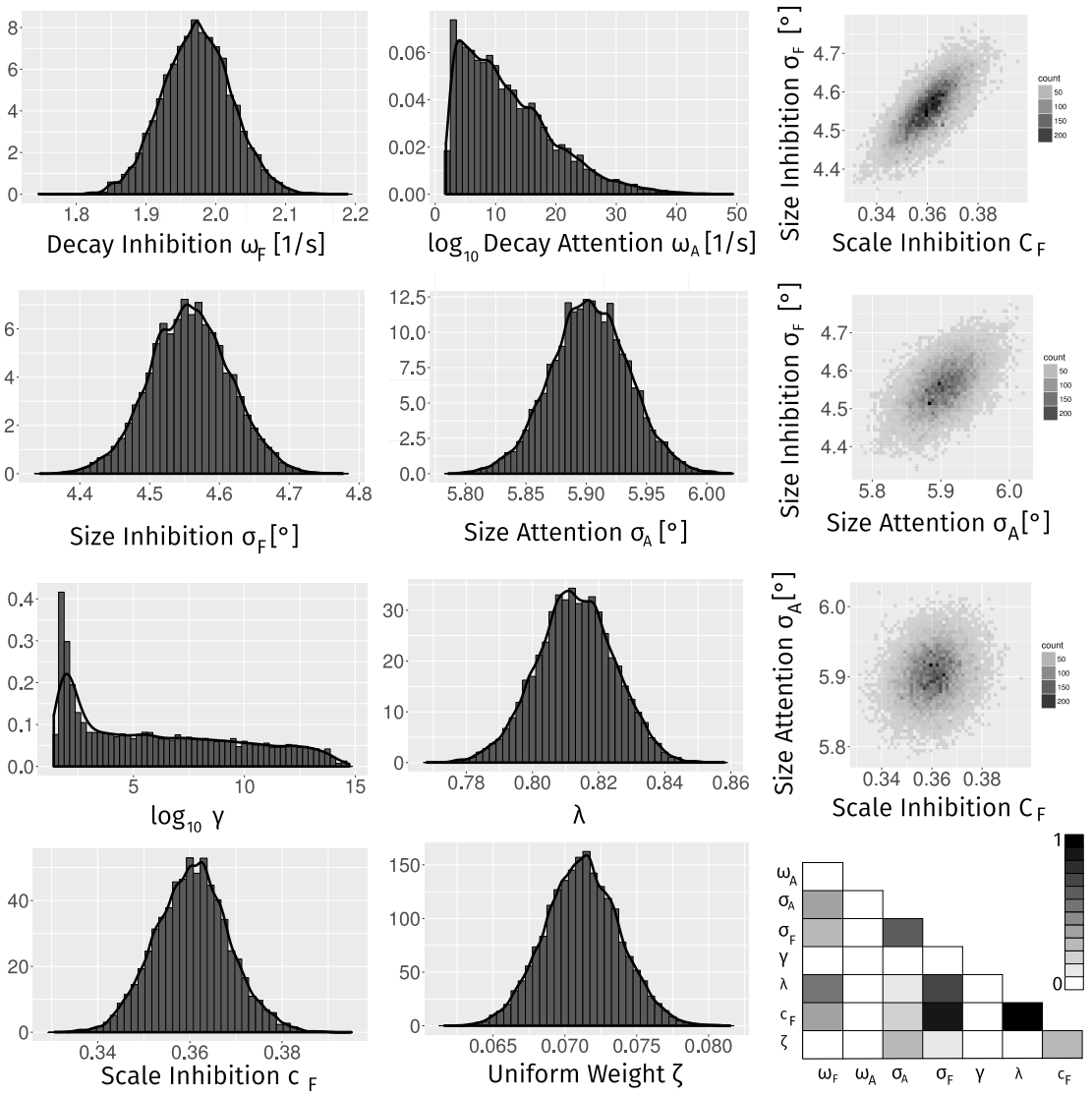

Figure 20: Sampling results for the posterior distribution for the example model's parameters. In the left two columns I show histograms and density estimates for all 8 parameters. Except for γ and ω_A all parameters seem to be well constrained by the data. In the right column I show two dimensional histograms of two parameters against each other illustrating their dependencies. The first indicates the strong correlation between the spatial scale and scaling factor of the inhibition. The second shows the medium strength dependency between the sizes of inhibition and attention pathway. The third plot illustrates the near independence of the spatial scale of the attention map and the scaling factor highlighting the non transitivity of correlations. In the lower right corner I present a summary plot about the correlations between parameters. The darkness of each rectangle in this plot indicates the absolute correlation between two parameters, which each could be shown as a 2D histogram as I did for 3 examples above.

Table 2: Table of the parameter values obtained from different point estimates. Displayed are the maximum likelihood estimate (MLE), the posterior mean estimate (± its estimated sampling error) and a credible interval from the Bayesian estimation I present, compared to the values from the original study by Engbert et al. (2015). Values marked with * were fixed without fitting in Engbert et al. (2015). The lower credible interval value for c_F printed in italics was reported wrongly in the article and is corrected here.

parameter	original estimate	MLE	posterior mean estimate		95% credible interval	
ω_A	6.607	2.4×10^{30}	1.1×10^{45}	$\pm 8 \times 10^{44}$	417.6	4.373×10^{30}
ω_F	0.00903	1.9298	1.973	±0.001601	1.876	2.071
σ_A	4.88	5.9082	5.903	±0.000640	5.838	5.967
σ_F	3.9436	4.5531	4.558	±0.002282	4.445	4.671
γ	0.3*	44.780	3.3×10^{12}	$\pm 4.5 \times 10^{11}$	43.83	3.249×10^{13}
λ	1*	0.8115	0.8130	±0.000422	0.7896	0.8354
c_F	1*	0.3637	0.3605	±0.000321	*0.3450*	0.3767
ζ	—	0.0722	0.0712	±0.000046	0.0662	0.0764

From an analysis of the marginal posterior distributions displayed in Figure 20, I can extract point estimates and credible intervals, which characterize a single optimal model parameter and a range that contains the true parameter value with a given probability. For the SceneWalk model I extracted the mean estimate and a 95% credible interval for each parameter listed in Table 2 to compare them to the parameter estimates obtained in the original paper (Engbert et al., 2015). For the well constrained parameters the MLE and mean estimates agree closely as expected. These estimates can only differ when the posterior is relatively broad. Consequently, our interpretation is the same for both parameter estimates.

Qualitatively, I reproduce the patterns observed in the original paper: The activation on the excitatory attention map is larger and faster than the inhibitory fixation map ($\omega_A > \omega_F$, $\sigma_A > \sigma_F$). Quantitatively, the parameters differ substantially from the ones in the original study. In particular, compared to the original study, (i) the Gaussian input around the current fixation is larger by roughly a degree for both maps, (ii) the inhibitory fixation map is 2.5 log-units faster, the attention map could be arbitrarily fast and (iii) the pooling exponents (γ and λ) converged to very different values than those chosen by hand.

The fact that the two parameters γ and ω_A are not well constrained can be explained as follows. The parameter ω_A determines the rise-rate of the attention map. Once this rate is fast enough, changes of the parameter value will not influence predictions any more. Similarly high values of gamma produce all very similar nonlinearities in the inhibition map and thus do not change any predictions. As I discussed above one could have used a prior to restrict these parameters to ranges over which they change predictions to avoid the result of parameters which are unconstrained over such wide ranges. This would however hide the fact that they are not well constrained from the posterior sampling result.

From the posterior distribution, I can also extract two-dimensional marginal distributions as histograms or density estimates. These marginal distributions illustrate posterior couplings between pairs of parameters. Such couplings indicate that obtaining information of one of the two parameters would constrain both of them better. For example, I show two-dimensional histograms for 3 pairs of parameters (Fig. 20):

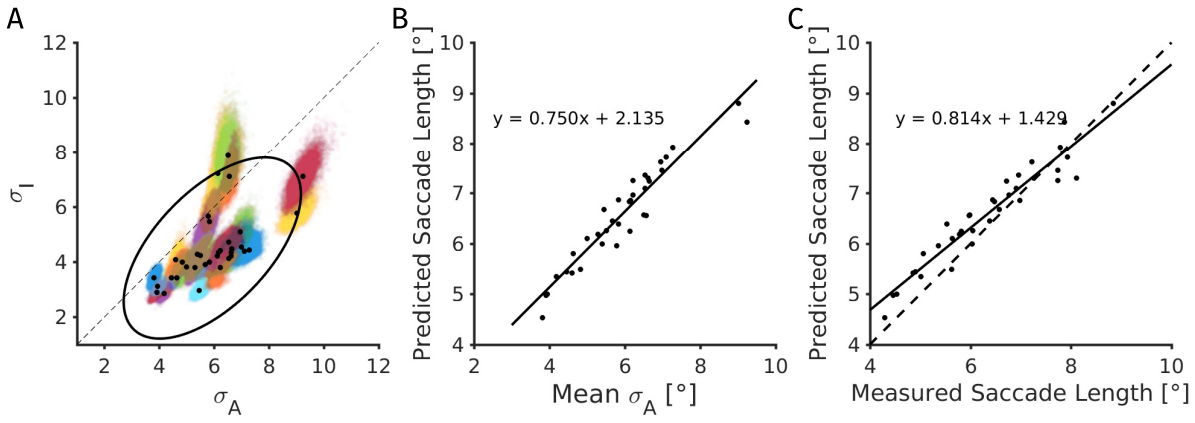

Figure 21: Results for the Hierarchical model. **A**: Fits for the two parameters σ_A and σ_F for the different observers. Each observer is represented by a black marker marking their posterior mean and a colored point cloud representing the posterior samples. Additionally the dashed line marks the $\sigma_A = \sigma_F$ diagonal and a large black ellipse marks the posterior mean estimate for the 95% line of the Gaussian population model. **B**: Predicted saccade length for each subject against their posterior mean estimate for σ_A with a linear least squares regression line. **C**: Predicted mean saccade length from the posterior mean estimate against the measured mean saccade length for each subject. The dashed and continous line mark the equallity diagonal and a linear least squares regression line.

- For σ_F and C_f I find a relatively strong coupling which indicates that models with stronger inhibition require it to be spread wider to explain the data equally well.
- For σ_A and σ_F I find a weaker, but still visible coupling, which indicates that the inhibition and attention window need to covary in size to explain the data.
- Finally, σ_A and C_F turned out to be approximately independent. Fixing one of these parameters would not constrain the other parameter.

This last point additionally illustrates that posterior correlations are not necessarily transitive.

In summary, the posterior marginal distributions can be reduced to the correlation coefficient, which captures the strength of the linear dependence between the parameters. These correlation coefficients are also plotted in Figure 20 for each combination of two parameters. The samples from the posterior also contain all higher-order dependencies between parameters, although they are more difficult to visualize or summarize.

3.3.4 *Inter-Subject differences and hierarchical models*

For many cognitive tasks subjects differ in meaningful ways, which we might want to include into our models. For eye movements, one important subject-specific parameter is the average length of saccades (Castelhano & Henderson, 2008). For our participants who generated the longest saccades, I observed average saccade lengths twice as large as the saccade lengths for participants with the shortest saccades (see Figure 21).

One popular method for integrating differences between subjects into models are hierarchical models. In hierarchical models the differences between subjects are explained by assuming different parameter values for each subject which follow an additional model

for the distribution of parameters in the population.[3] The main advantage of using a model for the distribution of parameters in the population is to stabilize the estimates for subjects, whose parameters are not well constrained by the data alone.

We implemented a hierarchical model which allows the sizes of the attention span and of the inhibited area to differ between subjects in order to explain the observed differences in saccade length. To simplify the analysis I fixed all other parameters of the model to their MAP estimates over all subjects and images from the model fitting explained above.

As our model for the parameter distribution in the population, I introduced a two dimensional Gaussian, which I parametrized using means and variances for the two parameters and the correlation between parameters as a fifth parameter. As I now aim to estimate these five parameters together with the individual subjects parameters, I defined a prior on each parameter individually and assumed the priors to be mutually independent. For each of the means and their correlation I chose a uniform distribution, while for the variances I selected an inverse Gamma distribution with parameters 0.25 and 1, which yields a very broad distribution over the positive real axis with a peak at 1.

It is possible to fit the hierarchical model using the same procedures I applied to the original model. I skip optimization and Frequentist analysis here though. Instead I directly sample the posterior using Gibbs sampling (Casella & George, 1992) with parameter groups for each subject and one group for the hyper parameters, sampling each marginal distribution using the Metropolis Hastings algorithm. Specifically, I first cycled through each subject performing one Metropolis Hastings sampling step for the corresponding two individual parameters. Next, I performed one Metropolis Hastings step for the parameters of the Gaussian distribution, which was assumed for the parameter distribution in the population. All proposal distributions were Gaussians with diagonal covariance matrix, adjusted by hand to approximately achieve 25% acceptance rate, and variances roughly proportional to the posterior variances of the parameters (Gelman et al., 1996). I used the same proposal distribution for each subject. Gibbs sampling is especially efficient for hierarchical models, since sampling the parameters of each subject requires only the likelihood for the data of that subject. Thus a whole sweep is computationally only as costly as single likelihood evaluation for updating all parameters. I sampled 3 chains of 10 000 sweeps through the parameters each starting at the maximum a posteriori estimates over all data. As burn in I removed the first 1 000 samples of each chain, which seemed sufficient after visual inspection of the chains. This yielded an effective sample size between 347 and 4472 for the different parameters and the chains seemed to have converged according to visual inspection of the chains and the $\hat{R}$ statistic which had an upper CI bound of 1.06 or less in all cases.

The results of the hierarchical model analysis are shown in Figure 21. First in A, we observe that different subjects are fitted by considerably different sizes for both σ_A and σ_F and that the estimates for the two parameters are highly correlated, i.e., subjects who have a larger fitted attention span also have a larger fitted inhibition area. Second in panel B I show that the mean saccade length predicted by the model depends strongly on σ_A and consequently on σ_F, as they are highly correlated. Finally I compare the measured mean saccade length to the mean saccade length predicted by the fitted model by simulating as much data as measured for each subject with their posterior mean parameters. The two observables are strongly related, indicating that varying the two spans in the SceneWalk model could account for the difference in saccade length between subjects. Additionally we can observe that the predicted mean saccade length grows with a slope slightly smaller

3 The hierarchical model framework can also be used to model effects of other properties of the task like item and image effects.

than 1 with the measured saccade length, indicating a slight regression to the mean, as expected and intended for a hierarchical model.

Looking at the individual subject estimates more closely, we can observe that most subjects (30 of 35) fall into a large cluster, with slightly smaller σ_F than σ_A. However, three subjects have larger fitted inhibition spans and two subjects have extraordinarily large attention and inhibition spans.

3.4 MODEL COMPARISON IN THE LIKELIHOOD APPROACH

The likelihood concept can be used as a general approach to evaluate how well a given model fits experimental data. Thus, it is possible to compare different models. For likelihood-based comparisons between models one usually assumes fitted parameters. Thus one uses the maximum likelihood, i.e., the best likelihood value a model can reach on the data, when the model's parameters are optimally adjusted. In the following, I denote the maximum likelihood as $L(M) = \max_{\boldsymbol{\theta}} L_M(\boldsymbol{\theta}|\text{data})$.

For the comparisons that I will carry out below, it is important that the log-likelihood is always a relative measure, since it depends on the grid for the observation of fixation positions, the size of the dataset and other dataset specific aspects. Therefore, only the log-likelihood-ratios between models can be compared between different datasets, models, or viewing conditions. Given a null model M_0, which defines a reference point, one can compute a likelihood ratio Λ to compare a model M_1 to the model M_0, i.e.,

$$\Lambda(M_1) = \frac{L(M_1)}{L(M_0)} \,. \tag{24}$$

The likelihood ratio Λ informs about how many times more likely the data are generated by model M_1 than by model M_0. For theoretical considerations and for most computations the log-likelihood ratio λ is a better choice,

$$\begin{aligned} \lambda(M_1) &= \log(\Lambda(M_1)) = \log \frac{L(M_1)}{L(M_0)} && (25) \\ &= \log(L(M_1)) - \log(L(M_0)) \,. && (26) \end{aligned}$$

The log-likelihood ratio is additive and can be interpreted in a straightforward way, e.g., if M_2 is one bit better than M_1, which is one bit better than M_0, then M_2 is two bits better than M_0 and the data are 4 times more likely under model M_2 than under model M_0.

Also, the log-likelihood ratio can be interpreted in information theoretic terms as the *information gain* about the data generated by the new model compared to the information explained by the original model. Thus the log-likelihood ratio measures how much communication could be saved when specifying a sequence of fixations using a code based on the model. As information theory is well developed (Ash, 1990, for an introduction), it provides a strong theoretical background for log-likelihood ratios in model comparisons.

In principle likelihood ratios measure the relative quality of the model fits. However, models tend to fit aspects of the data which are purely random, a phenomenon known as *overfitting* (e.g., Dietterich, 1995). Overfitting is the main reason why *model selection*—to which Zucchini (2000) gives an introduction for psychologists—should not be done by directly comparing the likelihoods based on the data used for fitting the models (Myung, 2000). Ultimately the goal of model comparison approaches is to compare the expected likelihood on new data, not on the data used for fitting. Proper model selection and

comparison methods are especially critical for comparing models which differ in their flexibility. More flexible models always explain more details of the dataset they are fit to, and thus produce larger likelihood values for the dataset they are fit to. However, more flexible models should only be preferred if the additionally explained details generalize to new data.

There are two popular quantities model comparison techniques try to estimate and use for comparing models. The first one is the *out-of-sample-prediction error* (Gelman, Hwang, & Vehtari, 2013), i.e. one tries to estimate the likelihood of the parameters fitted on the given data on a new dataset. The second one is the *evidence* for a model which is the denominator of the Bayesian formula—$\int_\Omega P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}$—i.e. the total probability to observe the data according to the model with the given prior $P(\boldsymbol{\theta})$. For a new dataset this means the evidence estimates the models performance using only the prior information about the parameter value. Consequently the evidence critically depends on the prior and can be arbitrarily bad if the prior assigns large probability to parameters with low likelihood. The ratio of evidences for two models is called the Bayes factor.

The first approach for model selection are metrics which add a correction or penalty term for more flexible models. These metrics are generally called information criteria and are usually formulated in terms of the *deviance* ($-2\lambda(M)$)—a general measure of prediction error—which is directly computed from the likelihood and contains exactly the same information, but reverses the sign. Thus smaller information criteria correspond to better models.

Classical examples for this procedure are the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). The AIC was formally introduced as a first model selection criterion, defined as: $AIC(M) = -2\lambda(M) + 2\dim(M)$[4]. It represents a simple large sample bias correction obtained from Fischer information theory estimating out-of-sample-prediction error. The BIC (Schwarz, 1978) was introduced as an approximation to the evidence in favour of a model in the case of an exponential family model. Thus it effectively aims to estimate the generalization quality to new data which requires new fitted parameters. For n independent observations it is defined as[5]: $BIC(M) = -2\lambda(M) + \log(n)\dim(M)$[4]. This obviously does not contain the prior and is a coarse approximation to the evidence. From very small datasets on this penalty will be larger for the BIC than for the AIC, e.g. the BIC will prefer parsimonious models more strongly than the AIC corresponding to the harder generalization task estimated by BIC.

The classical information criteria—AIC and BIC—both result in very small corrections of the raw likelihood. Our dataset contained 13908 and 13306 fixations for natural images and texture images respectively. Thus for our model with 8 free parameters the AIC and BIC penalties would maximally be $0.0008\frac{\text{bit}}{\text{fix}}$ and $0.0041\frac{\text{bit}}{\text{fix}}$ respectively, while the differences between models are much larger. In contrast, our cross validation results below suggest that the actual difference between fitted data and new data is much larger. Thus AIC and BIC seem to provide bad estimators in our case of complex dynamical models.

Very similar Bayesian evaluations exist (Gelfand & Dey, 1994), which estimate generalization of the posterior predictive distribution instead of generalizations based on a point

4 $\dim(M)$ representing the dimensionality of the model, i.e. the number of parameters, n the number of independent observations

5 The original criterion was half the value described here. However the version reported here seems to be the more commonly used one today.

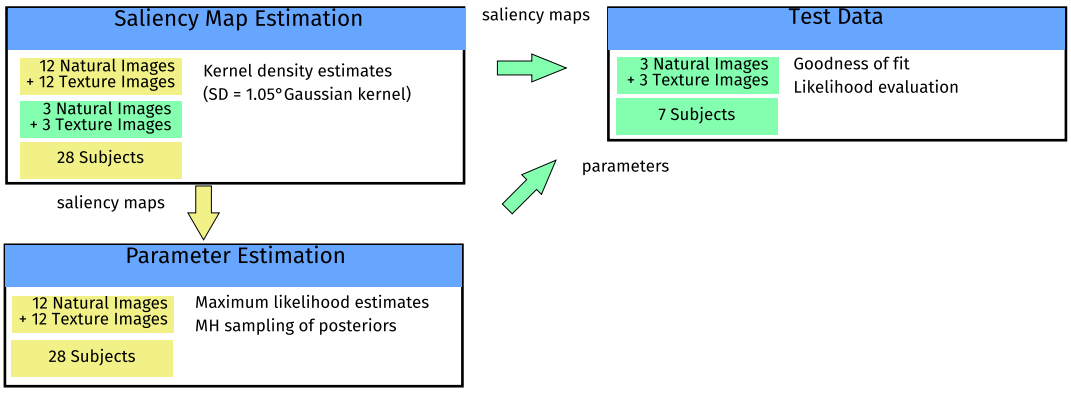

Figure 22: To guarantee that the model is fit to a different dataset than the one used for evaluation many possible separations exist. Here I display the separation of the dataset into training and test data used for each fold of cross validation. Data from 28 human observers on 2×12 images (yellow) were used for parameter fitting, while the data from 7 different observers on 2×3 test images were used for model tests (green).

estimate for the parameters. Nonetheless, the aim stays to predict how likely new data will be according to the model.

Fortunately direct formulas to approximate model performance in fully Bayesian terms from sampling results exist (Gelman et al., 2013). Thus a Bayesian Model comparison is possible, once a representative sampling is available for the posterior on the parameters of each model. Examples for this approach aimed at generalization to new data from the same parameters are the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) which approximates the posterior as the mean estimate and the Widely Applicable Information Criterion (WAIC, Watanabe, 2010), which directly uses the sampling estimate for the posterior predictive. Both these criteria also use the posterior samples to their advantage to produce a more accurate estimate for the out of sample prediction quality. Similarly, there is also a Bayesian alternative to the BIC, the Widely Applicable Bayesian Information Criterion (WBIC, Watanabe, 2013).

Calculation of the Bayesian information criteria requires an estimate for the posterior distribution on the model parameters, i.e., a sampling of the posterior. As I compare 10 models below and only have a sampling for one of these models, I do not perform these analyses here. However, such analyses should be considered especially when one studies other models like hierarchical models for example for which cross validation is not straight forward. And of course, once the posterior predictive is used for prediction, this should be the measure to be compared in the cross validation.

One should note that the penalties of all information criteria per data point (i.e., fixation or scanpath) converge to zero for growing dataset size. Thus larger datasets will raise a preference for more detailed models if there is any advantage for prediction. This makes sense as the criteria penalize complexity only when this complexities cannot be calibrated well enough to improve predictions with the given data (Burnham & Anderson, 2004).

A different more data driven approach to estimate the quality of out of sample predictions is *cross validation*, which is frequently used in machine learning, but has been introduced to the psychological literature as well (Browne, 2000). For cross validation the dataset is split into n subsets. Then the model is fitted to $n-1$ of the subsets—the *training set*—and evaluated on the one subset not used for fitting—the *test set*. This is

repeated for each of the subsets being the test set and the results are averaged. This procedure applies to Bayesian and Frequentist evaluation equally, but is more frequently used with point estimates and Frequentist evaluation.

For dynamical models for eye movements in scene viewing, two separate factors induce variability for which overfitting could occur: human observers (subjects) and stimuli. To avoid problems of overfitting for these two factors, I split the data across both factors and perform 5-fold cross validation using splits into training and test set as illustrated in Figure 22: For each fold I used the data obtained from 28 subjects on 12 natural images and 12 texture images for *training*. For evaluation I run the model on data obtained from 7 other subjects on 3 other natural images and 3 other texture images. To compute the empirical saliency maps, I used the 28 training subjects on both training and test images. There are also data for the training subjects on the test images and the test subjects on the training images, both of which are not used here to completely isolate training and test sets from each other.

For each fold I fitted the model to the training data using the genetic algorithm of MATLAB with settings as for the original fitting process on all data described above. However I noted that there was exactly one more local maximum to be found at small ($\sigma_F \approx .5°$), fast ($\omega_F \geq 10$) inhibitions, to which the genetic algorithm converged for some folds. To find the global maximum in every case nonetheless, I started a subsequent fminsearch optimization from each of these 2 maxima for each fold and took the better one as the global maximum. In all folds and all models the global maximum had similar sized attention window and inhibition and generally similar parameter values to the fit of the subtractive model to all data described above. The other local maximum was usually around 1000 worse on the log-likelihood scale for the training data. Thus the decision was always clear cut. Nonetheless this additional local maximum can be understood. Effectively it implements an inhibition for saccade targets very near to the current fixation. Saccades to these targets would not be detected as such by the data preprocessing such that such short saccades indeed do not occur in the dataset and cannot occur in a dataset. Thus this model adaptation indeed would be predictive, but not informative about any underlying processes of eye movement behaviour.

3.4.1 *Results on model comparison*

To perform the comparison I split the data as explained above, fitted the model to each of the 5 training sets and computed the log-likelihood of each model on each test dataset. Then I divided the resulting likelihood value by the number of fixations to normalize the results regarding the size of the dataset. Thus I measure all differences in bits per fixation [bit / fix]. My null model, the uniform distribution over the whole image, reaches a probability of 2^{-14} for every fixation to each grid point, since I calculated all maps on a 128×128 grid. This likelihood results in a log-likelihood of $-14\,\mathrm{bit/fix}$. I ran separate evaluations for texture images and object-based natural scenes presented in the experiments; the log-likelihoods are plotted in Figure 23. Overall, I find a gain for the empirical saliency model over center-bias prediction and a considerable gain in likelihood for the SceneWalk model.

The information gain for the saliency model differs strongly between natural textures and natural scenes, which was expected as the gaze patterns over texture images were more uniform than the corresponding data for natural scenes. This difference carries over to our dynamical model, as this uses the empirical saliency as an input predicting where human observers want to look. However, the increase in likelihood due to the

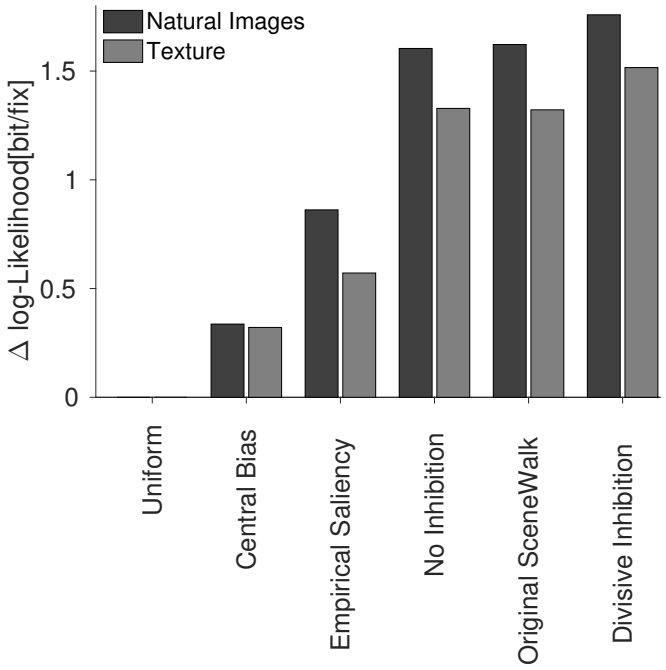

Figure 23: Bar plots for the models' log-likelihood differences to the uniform distribution null model. I split here by the two experimental conditions, which differed in the images presented. For the texture models the density map is much less informative than for the natural images. The *central bias/central fixation bias* model is a kernel density estimate from the fixations on all other images. The *empirical saliency* is the kernel density estimate from the fixations of other observers on the same image. Finally, *No inhibition*, *Original SceneWalk* and *Divisive Inhibition* refer to the three variants of the *SceneWalk* model, which I investigate in detail here.

dynamical principles is comparably large for texture images and for scenes. This result lends support to the view that the same dynamical principles of scanpath generation are underlying texture images and natural scenes.

We also evaluated the model with the parameters values fitted by Engbert et al. (2015). This yields a likelihood value of -12.96 bit/fix for natural images and -13.10 bit/fix for texture images for the training data (not shown in the figure). This indicates that the model explained the data better than empirical saliency even with the parameters not optimized for the likelihood. However, with the new parameter values the model generates higher likelihood values per fixation on the test sets it was not trained on (natural scenes: -12.38 bit/fix, textures: -12.68 bit/fix).

To compare different model specifications against each other, I generated two new model variants—one without inhibition and one with divisive inhibition—described in detail above. Additionally I questioned whether the introduction of the exponents λ and γ were necessary. To test this proposition I generated model variants with one or both of the exponents fixed yielding 4 variants of the subtractive original SceneWalk model, 4 for the divisive model and 2 for the model without inhibition.

First, as a check on the results it is informative to look at the performance of the models on the training data, I display in Figure 24A, although these values should not be used for model comparison. Evaluated on the training data a model which contains another model as a special case must be at least as good as the contained model on each of the training sets. This sanity check was how I first noticed that some of the optimizations

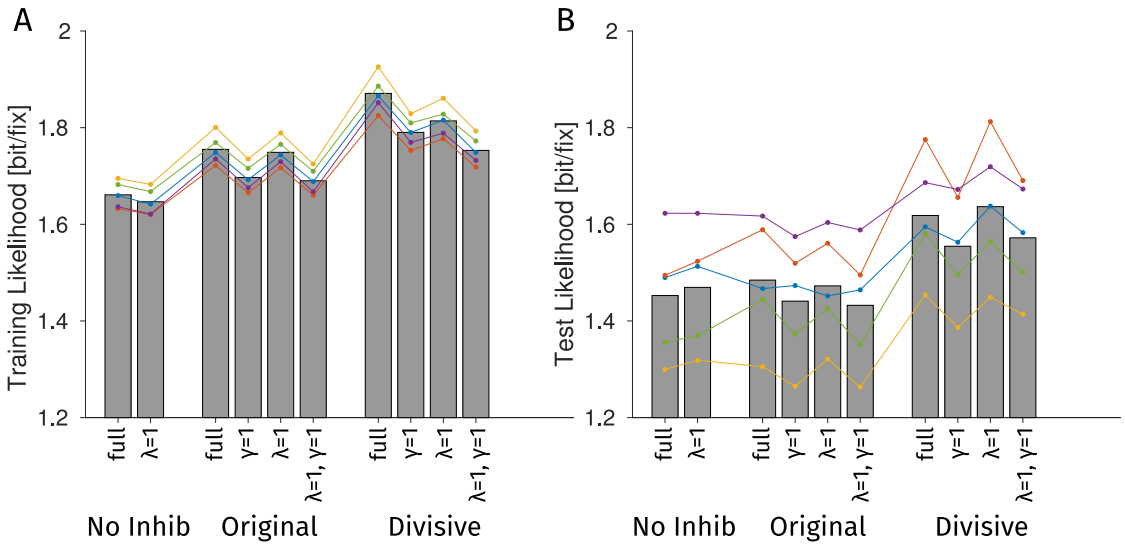

Figure 24: Bar plot comparing log-likelihood differences to the uniform distribution null model, exploring the effects of the exponents. Each bar is the average test set performance of the 5 folds of our cross validation procedure. The colored lines plot the results for the 5 folds. **A:** The likelihoods on the training datasets, which should not be used to judge the models, but are informative, whether the model fitting worked properly. **B:** The likelihoods on the test datasets, which can be used to compare models.

had ended in a different, wrong local maximum. Also comparing the training set and test set results provides some insight how substantial the flexibility problem is for the specific model.

The test set results of these more detailed comparisons are displayed in Figure 24B. I find that the divisive inhibition model overall provides the best performance followed by the original SceneWalk model and finally the model without inhibition. Within each model type the exponent γ seems to improve the model fit, while the fits with free λ yield equally good performance or even worse performance than fixing $\lambda = 1$ (using the attention map without non-linear distortion). The model to choose from our pool is thus the divisive inhibition model with a large, fitted γ and λ fixed to 1.

Note that all the models with inhibition have a qualitatively similar behaviour and typically computed statistics on scanpaths cannot discriminate these models, as I discuss below. Thus the likelihood based comparisons allow us to differentiate models could not be differentiated otherwise. A restriction of these model comparisons is, however, that they do not come with a measure of uncertainty like standard errors, credible or confidence intervals or adequate statistical tests[6]. Thus we cannot provide a hard statistical measure how sure we are about the order of the models although the differences can be interpreted in size.

3.5 GOODNESS-OF-FIT FOR SPECIFIC MEASURES AND SPATIAL STATISTICS

While we used the likelihood as a general measure of model fit to experimental data, the likelihood remains a relative (i.e., depending on a null model) and global measure (i.e., no specific statistical properties are addressed). Thus, there are at least two reasons to

6 Some classical χ^2 tests of model fit exist. As they are based on the same approximations as the AIC and BIC, I doubt that they produce correct conclusions here.

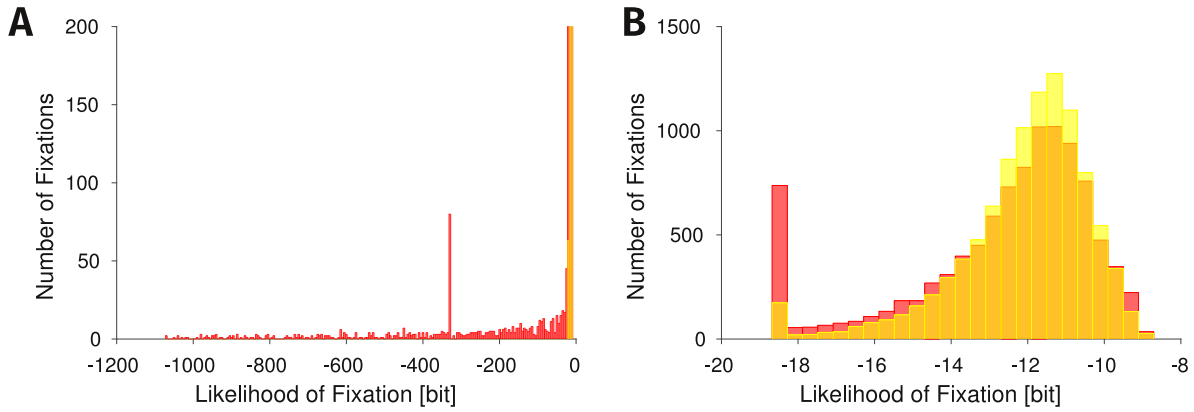

Figure 25: Histograms of the likelihood of individual fixations on the test dataset (red) and on data generated from the model (yellow) **A**: Employing a model without mixing with a uniform distribution (setting $\zeta = 0$ in Eq. (21)). The considerable number of extremely unlikely fixations led us to include the mixture with a uniform distribution in Eq. (21). **B**: Employing the full model with the mixture, extremely unlikely fixation positions no longer occur.

check other statistics additional to performing a likelihood-based approach to parameter estimation or model comparison. First, to analyze the absolute performance of the model, and, second, to understand which aspects of the data are modeled adequately and which other aspects are modeled poorly.

The first reason, judging the absolute quality of models, is to check that they are good enough to be interesting, which is subsumed under *goodness-of-fit* analysis in statistics (Pitt, Myung, & Zhang, 2002; Wichmann & Hill, 2001). In statistics, the importance of goodness-of-fit analyses is emphasized, since the theory of parameter estimation for models is built on the assumption that there is a correct solution, i.e., model parameter values exist that actually generated the data. So, if a model cannot explain the data well for any parameter value, the best estimate for the parameter might be meaningless, even when the best parameter value is defined by generating the highest likelihood for a given model. For the same reason, Bayesian inference methods may fail if there are no good models in the set assumed a priori.

To get an idea about the absolute quality of the model's predictions for data, the easiest way is to simulate data by the model and to compute statistics for these data in exactly the same way as it is done for the interpretation and statistical analysis of experimental data. A comparison of the resulting statistics gives a good indication of the quality of the model's fitness.

Based on the likelihood it is also possible to test how (un-)likely the measured data are, compared to the expected likelihood of data from the model. This expected likelihood can be computed by simulating larger amounts of data from the model and computing its likelihood. For a perfect fit, the measured data should have a similar likelihood as datasets simulated from the model, which represents a test whether the model's output variability matches the variability of the observed data.

We performed such an analysis by simulating as much data as we had collected and computed the likelihood of this data. We compare histograms over the log-likelihood per fixation for simulated and experimental data in Figure 25. First, in Figure 25A, we ran the analysis on a model without the mixture with a uniform distribution, i.e., choosing $\zeta = 0$. According to this model some of the observed fixations were extremely unlikely, i.e.

the model predictions were to specific, which motivated us to include the mixture with a uniform distribution. In Figure 25B, we show a histogram of the log-likelihoods for the full model, again for the measured data and simulated data from the model. For the full model, the mean log-likelihood of the simulated data is -12.11 bit/fix, $\Delta = 1.89$ bit/fix (raw value, difference Δ to a uniform distribution), which is roughly equal to the likelihood for the training data of -12.08 bit/fix, $\Delta = 1.92$ bit/fix, but larger than for the test data for which the model reaches only -12.67 bit/fix, $\Delta = 1.33$ bit/fix. The small difference between training data and model-generated data suggests that the model did not overfit the data dramatically, i.e., we would expect the model to be roughly as good as it is for the data, if the data were generated by the model. The difference between training and test data suggests that the model does not generalize to the test dataset perfectly, which is mainly caused by an increased number of highly unlikely fixations (Fig. 25B). It seems plausible that these are fixations in regions where none of the observers in the training set fixated (regions of low empirical saliency). This indicates that a higher number of observers for estimating the empirical saliency map would be beneficial to our approach.

The second motivation for additional model analyses is to decide which aspects of the data are modelled well, and which are not described adequately. It is important to further improve models and to choose appropriate models for different situations and modelling goals. Generally, measures used for this analysis should be interpretable for the modeller and other researchers. Some more detailed information can also be extracted from the likelihood calculations as this calculation is split over the different observations. Thus for each individual observation a separate likelihood can be computed and one can check which measured scanpaths or individual fixations are especially likely or unlikely according to the model providing some additional, more specific information.

For the SceneWalk model we started with an analysis of standard statistics from eye-movement experiments. As a first step, we compared the overall fixation density of model and data. To quantify the comparison, we computed the *Kullback Leibler Divergence* (KL-divergence) of the fixations predicted by the model against the fixations made in our experiment. This standard measure is computed as

$$KL = \int_I p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \ , \tag{27}$$

where the integral is computed over the full image I.

The fixation density generated by the model does not fit the empirical saliency perfectly, but perturbs it slightly through its dynamics. However, the predicted distributions diverge less from the true density (average KL-divergence = 0.1997) than any saliency models, which minimally reach 0.54 and 0.37 for the two datasets in the MIT saliency benchmark (Bylinskii et al., 2016). The good performance of the SceneWalk model is not surprising here, since we used the empirical fixation density as an input to our model.

Next, we looked at the distribution of the saccade lengths, a first aspect of the model dynamics. The results of this analysis are given in Figure 26. The saccade lengths in the model and data are very similar and the variance over images is small in both model and data, while the variance over subjects is substantial as we discussed above. Also the competitor models without inhibition and with divisive inhibition fit the distribution of saccade lengths well such that the saccade length distribution does not clearly differentiate these models from each other. However, simply drawing fixations independently from the empirical saliency map yields an entirely different, wrong distribution.

Recently, methods from the theory of spatial point processes were introduced into the analysis of fixation patterns in scene viewing (Barthelmé et al., 2013; Engbert et

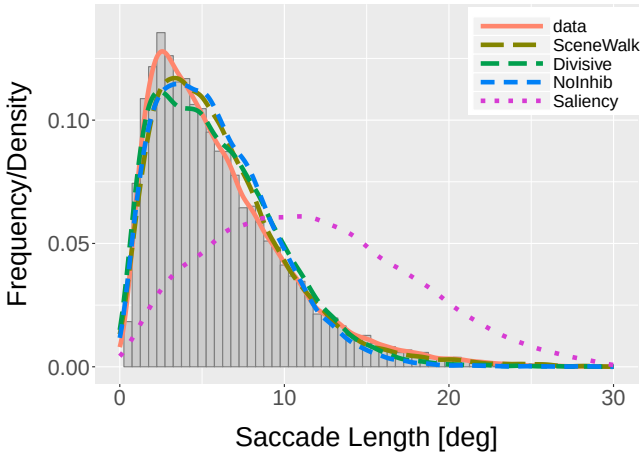

Figure 26: Comparison of model and data based on saccade lengths. The plots present the saccade length distribution over all images for experimental data and model simulations.

al., 2015). Most of the standard statistical measures are *first-order statistics*, e.g., the 2D density of fixations. For the SceneWalk model, we computed the *pair correlation function* (Engbert et al., 2015) as an example for a second-order spatial statistic. The pair correlation function (pcf) describes how frequently two fixations with a certain distance occur in one scanpath normalized against the frequency expected for a random selection from the fixation density. Values higher than one indicate that fixation patterns are more aggregated than could be expected from the first-order spatial inhomogeneity of the process. As the pair correlation function includes later returns to earlier fixated positions, this function measures a different property than the saccade length distribution. In experimental data, the pair correlation function usually indicates a clustering at small distances below $3 - 4^\circ$ (Engbert et al., 2015). Comparing the pair correlation functions estimated from experimental data and model predictions in Figure 27, it is obvious that all models fit the pair correlation function much better than a simple random process that draws fixations from the empirical density map. However this measure seems not to differentiate between the different types of inhibition either.

3.6 DISCUSSION

The key motivation for the current study was to apply the likelihood approach to the evaluation of dynamical cognitive models and, in particular, for model parameter estimation and model comparison. Dynamical cognitive models are formulated by evolution equations (temporally discrete or continuous) and evaluated against time-ordered data (time series). As a specific example, we investigated the problem of saccade generation, where the dynamical model determines the probability $\pi(x, t)$ to select a saccade target position x at time t. In the SceneWalk model (Engbert et al., 2015), this probability is computed from activation fields at any point in time. Thus, we can compute the corresponding probability for a fixation and force the model to generate the gaze shift to the new fixation position. This procedure of direct computation of the likelihood will work for the broad class of dynamical models that generate continuous-time activations for the prediction of discrete behavioral events (Erlhagen & Schöner, 2002).

For the interpretation, we normalized the likelihood with respect to the number of fixations in a given dataset to obtain a measure that is independent of the size (length) of the fixation sequence. Furthermore, we suggested to compare the likelihood to the

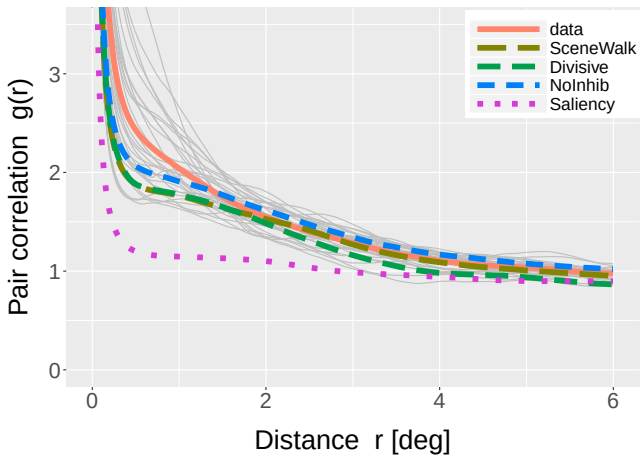

Figure 27: Comparison of models and data based on the pair correlation function (PCF). The mean PCF for each of the models is plotted in color. For the data the mean is shown in color as well and the pair correlation functions for individual images are plotted in gray. Higher values than one indicate clustering or aggregation, i.e., fixations at distance r are more abundant than expected on average from independently drawn fixations from the fixation density. Values smaller than one indicate repulsion, i.e, fixations at distance r are rarer than expected for independently drawn fixations.

likelihood obtained from a uniform distribution to get a measure which is independent of grid and image sizes. For simpler, non-dynamical models this comparison to chance performance is a standard procedure. Additional non-dynamical models were used to generate likelihoods to compare to the dynamical model. Such non-dynamical density models (e.g., the central fixation bias, Tatler, 2007) represented a convenient statistical baseline for our computations. Finally, we investigated two variants of the SceneWalk model to show that the likelihood can be applied as a powerful tool to distinguish different dynamical models with highly specific assumptions.

The likelihood as a global measure of model performance can be used as a tool for the estimation of model parameters. Fitting models based on the maximum likelihood concept has a long tradition in statistics and some clear advantages over other parameter fitting procedures, including mathematical proofs for the convergence and sufficiency of the parameter estimate. A practical advantage is that the likelihood is a scalar value, which does not rely on simulating complex discriminating statistics. Additionally, model fitting based on the likelihood is the starting point for Bayesian inference about parameter values, which provides new insights to other parameters that could explain the data and, thus, statistical comparisons on whether the parameters differ between datasets or conditions.

For the SceneWalk model (Engbert et al., 2015) we computed parameter values using maximum likelihood estimation and sampled the posterior for Bayesian parameter estimation. This parameter estimation technique allowed us to fit all the parameters of the model, which was impossible in the original publication. The parameters found by optimizing the likelihood reproduce all the statistics the original publication reported, while the parameters from the original publication perform significantly worse in terms of likelihoods. Additionally, we computed a full posterior probability over the parameters that informs about which parameters are constrained by the data well and which parameters are not constrained by the data.

Furthermore, the likelihood-based evaluation helped us to improve the original model. Using a hierarchical model, we found that the known differences between subjects in

their average saccade length (Castelhano & Henderson, 2008) could be fit well, by allowing the size of the attention window and the size of the inhibition to vary between subjects. Furthermore likelihood based comparisons between models allowed us to show that the dynamics and the inhibition both improve model predictions. And additionally we could differentiate different variants how the excitatory and inhibitory maps are combined. For experimentally-motivated statistics, these specific model variants made very similar predictions. Among the models analyzed here, a divisive inhibition model with a fixed numerator exponent λ seems to fit the data best—and even better than the original SceneWalk model.

With the SceneWalk model, we focus on fixation locations and take fixation durations as given (or a random process with given mean and variance). This is, however, not necessarily a restriction of the likelihood approach. Models which compute probabilities for fixation durations (Nuthmann et al., 2010; Trukenbrod & Engbert, 2014, for example) or for both the durations and locations of fixations could be fit and evaluated using the same techniques we present here for locations only. There are recent studies on fixation durations for scene viewing (e.g., Laubrock, Cajar, & Engbert, 2013). Furthermore, the prediction of fixation durations is a main aim for models of eye movements during reading (Engbert et al., 2005; Reichle et al., 2003).

In this article we used relatively simple gradient free optimization algorithms and the Metropolis-Hastings algorithm for their conceptual simplicity, which eased the presentation. However, there might be more efficient algorithms for solving the optimization and sampling problems in the SceneWalk model and certainly different algorithms will be best or easiest to implement for different models. Also, the optimizations and samplings for complex models may take hours, days or even months of computation time. Thus efficiency is important as it may make the difference whether an analysis is feasible with given computational resources or not. Consequently, it can be worthwhile to invest some time to try different optimization algorithms including global optimization algorithms, when local minima are a problem. Similarly there is broad literature on how to (adaptively) tune MCMC-algorithms (e.g., Andrieu & Thoms, 2008; Gelman et al., 1996; Haario, Laine, Mira, & Saksman, 2006; Haario, Saksman, & Tamminen, 2001; Roberts & Rosenthal, 2009) and efficient sampling algorithms (Brooks et al., 2011; Robert & Casella, 2009, 2013).

An especially large step in efficiency for both optimization and sampling can be made if a gradient of the likelihood can be calculated with reasonable efficiency. For optimization highly efficient gradient based algorithms, i.e. quasi-Newton methods like the BFGS algorithm are available. The original gradient based sampling algorithm is the Hamiltonian Monte Carlo (HMC) method introduced by Duane, Kennedy, Pendleton, and Roweth (1987) (see Neal, 2011, for an introduction). By now there are many variants of HMC available, including adaptive methods like the No-U-turn Sampler (NUTS, Hoffman & Gelman, 2014), which works behind STAN (B. Carpenter et al., 2017), one of the most recent general purpose samplers. These samplers contain automatic differentiation tools, which remove the necessity to code a gradient computation by hand. Also independent tools to compute derivatives automatically are able to differentiate virtually any computable function (Abadi et al., 2015; Theano Development Team, 2016), which allows computation of a derivative for many models.

As a next step the likelihood evaluation permits comparisons between different models. To avoid overfitting such comparisons were carried out using cross validation. Here, the SceneWalk model (Engbert et al., 2015) was compared to a statistical model of the central fixation bias and to a model that sampled fixation positions from the empirical saliency

map. We found that the SceneWalk model outperforms the empirical saliency model by $0.75\frac{\text{bit}}{\text{fix}}$, which highlights the importance of incorporating influences of previous fixations into predictions for upcoming saccade targets. Consequently, a saliency model alone is not a good models for scanpaths, no matter how closely it matches the fixation density.

As the likelihood is a relative measure, it is necessary to check whether the fitted model is reasonably good in terms of absolute measures. For the SceneWalk model we demonstrated the adequacy by comparing different summary statistics computed on model predictions to the corresponding statistics obtained from experimental data. We found that the model reproduced the fixation density, saccade length distribution and the pair correlation function with parameters computed via maximum likelihood estimation.

For scanpath models in eye-movement research, the likelihood approach to parameter estimation and model comparison is most interesting as there is no general consensus on a metric for comparing models so far (Le Meur & Baccino, 2013; Pitt et al., 2002). Instead, many statistics on specific aspects of scanpaths were proposed, which allow judgements whether a given model shows some specific effects or not. However, a global account of how adequately the model fits the experimental data is currently lacking. We demonstrated that such global measures could be provided by the likelihood approach.

In the likelihood approach, any scanpath observed in humans must have a probability larger than zero under the model, as the likelihood vanishes otherwise, indicating only that the model cannot explain the data. A second constraint on the model is that the likelihood can be computed. As we showed above, it is sufficient to be able to numerically generate the probability for the next fixation given the previous ones. This is not a strong constraint as most eye movement models on natural scenes even explicitly represent a probability map for the next fixation (Le Meur & Liu, 2015; Zelinsky, 2008; Zelinsky, Adeli, Peng, & Samaras, 2013, for example).

We believe that model evaluations based on the likelihood are promising for many other psychological models. Indeed, for some models the evaluation is already routinely done using likelihoods, for example for receiver operating curves (Ogilvie & Creelman, 1968), diffusion models (Ratcliff & Tuerlinckx, 2002) or psychometric functions (Wichmann & Hill, 2001) and recently for saliency models and fixations on static images (Barthelmé et al., 2013; Kümmerer et al., 2015).

One favourable aspect of the SceneWalk model is that it is deterministic—there is only a single way for the model to produce time-dependent activation maps for a given sequence of fixations. If there were multiple possible internal states compatible with the observed data, then the computation of the likelihood would require an integration over all possible internal states. Such integration could render evaluations of the likelihood function less effective or even impossible for other models. For such complex models with many possible internal states and large datasets efficient computational techniques for combined state and parameter estimation have been developed in particular in the field of *data assimilation* (Law, Stuart, & Zygalakis, 2015; Reich & Cotter, 2015). Furthermore, processing time-ordered datasets leads naturally to the consideration of *sequential Monte Carlo methods* (Chopin, Jacob, & Papaspiliopoulos, 2013; Doucet, de Freitas, & (eds.), 2001), to bring computational demands into a manageable range.

For some model classes computation of the likelihood might be too time consuming or the likelihood function too complex for further handling. However, even for such models, mathematically well founded approximations to the likelihood methods were proposed: *Pseudo-likelihood* methods compute an approximation to the likelihood (Wood, 2010, for example). Alternatively, *pseudo-marginal Monte Carlo methods* (Andrieu & Roberts, 2009; Beaumont, 2003) can be utilized which, while involving approximations,

can be shown to provide consistent estimates. Here one could also consider replacing the likelihood by an appropriate *scoring function* (Gneiting, Balabdaoui, & Raftery, 2007) which provides an alternative metric to rank models in an objective manner. Moreover, *Approximate Bayesian Computation* (ABC) allows an approximation to full Bayesian inference without a likelihood (Barthelmé & Chopin, 2011, 2014; Turner & Van Zandt, 2012; Wilkinson, 2013). These methods preserve some of the benefits of the likelihood approach to parameter estimation and model analysis and can even be used to do model selection. For dynamical models this is discussed for example by Toni, Welch, Strelkowa, Ipsen, and Stumpf (2009).

3.7 CONCLUSION

We proposed and studied a likelihood approach for the evaluation of a dynamical cognitive model for the control of saccadic eye movements. The likelihood can be used for parameter estimation and model comparisons as it makes the full range of statistics available, from maximum likelihood estimation through Bayesian estimation and hierarchical models to proper model comparisons. Compared to non-dynamical models, the dynamical model generated a significant increase in predictive power by introducing sequential dependencies. Our approach is a promising tool for the evaluation of dynamical models that predict sequences of discrete behavior (e.g., fixation position, movement onsets) in general and for human scanpaths in particular.

4

CONNECTING EARLY VISION AND EYE MOVEMENTS

Any picture (unless it is a uniform background or a repetitive mosaic) contains different elements; the eye rests much longer on some of these than on others. while some elements may receive little or no attention. What distinguishes the elements particularly attracting the observer's attention. and what are the characteristic features of those elements which do not draw his attention?

Yarbus (1967)

In this chapter I describe the implementation of a saliency model which predicts fixation locations based on the internal representations of the early spatial vision model I describe in Chapter 2 and evaluate it based on the methods discussed in Chapter 3. I evaluate this model, several other saliency models and the maximally attainable prediction performance over time to disentangle low- and high-level contributions to eye movement control. Furthermore I evaluate on a search dataset to disentangle top-down and bottom-up control.

The content of this chapter reflects the current state of a manuscript in preparation. Consequently, an article with similar contents to this chapter may appear soon. A summary of my contributions to this manuscript is given in a declaration accompanying this thesis as for the published articles.

4.1 INTRODUCTION

The guidance of eye movements in natural environments is extremely important for our perception of the world surrounding us. Visual perception deteriorates quickly away from the gaze position such that many tasks are hard or impossible to perform without looking at the objects of interest (reviewed by Strasburger et al., 2011, section 6; see also: Land et al. (1999)). Consequentially, the selection of fixation locations is of great interest for vision researchers and many theories were developed to explain the selection of fixation locations.

Classically, the factors determining the eye movements of human observers are divided into bottom-up and top-down influences (Hallett, 1978; Tatler & Vincent, 2008). Bottom-up influences refer to stimulus parts which independently attract fixations. The existence of bottom-up guidance of eye movements was originally postulated, because some stimuli like flashing lights made subjects move their eyes towards them under well controlled laboratory conditions, even in tasks when they were explicitly asked not to look at the stimulus, as in anti-saccade tasks for example (Hallett, 1978; C. Klein & Foerster, 2001; Mokler & Fischer, 1999; Munoz & Everling, 2004). How important bottom-up effects are under more natural conditions and especially for static stimuli remains a matter of debate. Top-down influences on the other hand refer to cognitive influences on the chosen fixation locations, based on the current aims of the observer like social implications (Emery, 2000; Friesen & Kingstone, 1998) or tasks and memory (Henderson et al., 2007; Land et al., 1999). The main argument for the involvement of top-down control are task effects on the fixated locations (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007; Underwood, Foulsham, Loon, Humphreys, & Bloyce, 2006). More recently systematic tendencies were introduced as a third category (Tatler & Vincent, 2008), which encompasses the dependencies between fixations like the preference for some saccade directions (Foulsham et al., 2008) or the dependencies between successive saccades (Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2016; Tatler & Vincent, 2008; Wilming et al., 2013). While all these aspects seem to contribute to eye movement control, the debate, how these aspects are combined and how important the different aspects are, continues till today (Borji & Itti, 2013; Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2008; Hallett, 1978; Harel et al., 2006; Kienzle et al., 2009; Schomaker, Walper, Wittmann, & Einhäuser, 2017; Stoll et al., 2015; Tatler, Hayhoe, Land, & Ballard, 2011; Tatler & Vincent, 2009).

In this debate the two sides typically argued either for top-down control based on high-level features (Castelhano et al., 2009; Yarbus, 1967) or for bottom-up control based on low-level features (Itti & Koch, 2001; Kienzle et al., 2009). These stances couple the question how complex the features for eye movement control are to the question how

much voluntary control we have over our eye movements. These questions are orthogonal however and the less usual stances are sensible as well. Bottom-up control may encompass not only low level features like contrast, colour or edges (Itti & Koch, 2001; Itti et al., 1998; Treisman & Gelade, 1980), but also high level properties of the explored scene like object locations (Einhäuser, Spain, & Perona, 2008) faces (Judd et al., 2009; Kümmerer, Wallis, & Bethge, 2016) or even locations which are interesting or unexpected in a scene category (Henderson et al., 1999; Torralba et al., 2006). This stance is embraced by most modern models for the prediction of fixation locations in images, which almost all use high level features computed from the image (Bylinskii et al., 2016; Judd et al., 2009; Kümmerer et al., 2016). Similarly, top-down control may not only act on high-level features but could also act on low-level features like contrast, orientations or colour. Such influences are especially important in models of visual attention (Müller & Krummenacher, 2006; Tsotsos et al., 1995) and of visual search (Wolfe, 1994), which often postulate top-down control over which low-level features shall guide attention and eye movements or even top-down influences on the processing of the low-level features.

One especially unclear term in this debate is saliency (Tatler et al., 2011), which originally stems from attention research, where it refers to objects which stand out from the other objects in the display and attract attention (Koch & Ullman, 1985). As the first computable models for the prediction of fixation locations in images were based on these ideas saliency was soon associated with these models and became synonymous with bottom-up low-level control of eye movements (Itti & Koch, 2001; Itti et al., 1998). As it became clear that the prediction of fixation locations requires high-level features they were added to the models, but the models were still referred to as saliency models (Borji & Itti, 2013; Bylinskii et al., 2016; Judd et al., 2009). Consequentially saliency in computer vision now refers to any image-based prediction which locations are likely to be fixated by subjects. To avoid the confusion associated with the term saliency I shall refrain from using it alone and will only use the combination "saliency model", which shall refer to any model which predicts fixation locations based on an image.

How the fixation density changes during the exploration and how the influence of different factors varies over time has been largely ignored, although this information might be important to understand the interplay of the different factors. This left a gap in our understanding I want to fill with this chapter. To do so, I employ the Corpus and the search dataset collected in Postsam recently which each contain exceptionally many fixations per image, such that I have enough data to sensibly estimate the fixation density at different points in time during the exploration of an image. In the new *Corpus* dataset 105 subjects tried to memorize the images, in the *Search* dataset (Rothkegel et al., 2018) 10 subjects searched for overlaid targets 48 times per image each.

Using these new data I want to address two overarching questions: First, how well can we predict the fixation density from the image in principle, i.e. what is the limit for bottom-up models and how does this aim change over time? Second, when is the fixation density more determined by low-level or high-level features?

To answer the first question I explore the performance of the empirical density over time. To quantify its performance I use the likelihood based techniques I presented in Chapter 3. These methods are especially useful here to avoid the ambiguities of using typical saliency model evaluation criteria and allow a unified metric for all models.

To answer the second question I additionally need to measure the influence of low- and high-level features. To measure the influence of low-level features one could choose from a range of classical low-level models (e.g. Itti et al., 1998; Kienzle et al., 2009), which certainly predict free viewing fixations above chance level and were originally claimed to

be good models of fixation selection in general (Itti et al., 2000, 1998). However, the features the low-level models use were only informally linked to the low-level features used in models of perception. To remove this ambiguity of interpretation I will present a new saliency model below, which is based on the representation produced by the model of early spatial vision I present in Chapter 2.

To argue for the influence of high-level features, earlier approaches made predictions based on object locations which predict eye movements better than the low-level saliency models of their time (Einhäuser, Spain, & Perona, 2008; Stoll et al., 2015; Torralba et al., 2006), experimentally varied low-level features (Anderson, Ort, Kruijne, Meeter, & Donk, 2015; Açık et al., 2009; Stoll et al., 2015) or chose specific examples for which low-level and high-level features make opposing predictions (Vincent, Baddeley, Correani, Troscianko, & Leonards, 2009). These classical approaches do not easily make predictions for new images however. Fortunately, the idea of object-based saliency map models could recently be unified with low level factors due to the advent of deep neural network models (DNNs; see Kriegeskorte, 2015, for an overview). These DNNs contain activation maps which effectively encode what kind of object can be found where in an image. Like simple low level features, these object based features can be used to predict fixation locations (Huang, Shen, Boix, & Zhao, 2015; Kruthiventi, Ayush, & Babu, 2015; Kümmerer et al., 2016; Pan et al., 2017). Saliency models based on this principle currently dominate the benchmarks (Bylinskii et al., 2016). Thus these DNN-based saliency models provide a better and more convenient quantification of what information can be predicted using high-level features than earlier approaches. As a representative I use the best performing of these models, DeepGaze II (Kümmerer et al., 2016).

I separate the presentation of the results by the dataset they are based on. I first develop our early vision based saliency model and evaluate it alongside the other saliency models over time on our corpus dataset. Then I analyse the saliency predictions similarly on the search data and end with the analyses specific to the search dataset.

4.2 METHODS

4.2.1 *Stimulus presentation*

Sets of 90 (corpus) and 25 (search) images were presented on a 20-inch CRT monitor (Mitsubishi Diamond Pro 2070; frame rate 120 HZ, resolution 1280×1024 pixels; Mitsubishi Electric Corporation, Tokyo, Japan). All stimuli had a size of 1200×960 pixels. For the presentation during the experiment, images were displayed in the center of the screen with gray borders extending 32 pixels to the top/bottom and 40 pixels to the left/right of the image. The images covered 31.1° of visual angle in the horizontal and 24.9° in the vertical dimension.

4.2.2 *Measurement of eye movements*

Participants were instructed to position their heads on a chin rest in front of a computer screen at a viewing distance of 70 cm. Eye movements were recorded binocularly using an Eyelink 1000 video-based-eyetracker (SR-Research, Osgoode/ON, Canada) with a sampling rate of 1000 Hz.

For saccade detection a velocity-based algorithm was applied (Engbert & Kliegl, 2003; Engbert & Mergenthaler, 2006). This algorithm marks an event as a saccade if it has a

Corpus Dataset

Figure 28: Overview over datasets. Left: Scene viewing Corpus dataset, for which eye movements of 105 subjects on the same 90 images were recorded with slightly varying viewing conditions asking them to remember which images they had seen for a subsequent test. Right: Visual Search task. Here eye movements of 10 subjects were recorded while they searched for the 6 targets displayed below the image for eight sessions each. In the experiment each image contained only one target and subjects usually knew which one. Additionally I increased the size and contrast of the targets for this illustration image to compensate for the smaller size of the image. The right panel is reused with permission from our article on the search dataset (Rothkegel et al., 2018).

minimum amplitude of 0.5° and exceeds the average velocity during a trial by 6 median-based standard deviations for at least 6 data samples (6 ms). The epoch between two subsequent saccades is defined as a fixation. All fixations with a duration of less than 50 ms were removed for further analysis since these are most probably glissades, i.e. part of the saccade (Nyström & Holmqvist, 2010). The number of fixations for further analyses was 312267 in the corpus experiment and 176 828 in the search experiment.

For calibration a 9 point calibration was performed in the beginning of each session of the Corpus-Experiment and of each Block of the Search experiment and revalidated the setup whenever the fixation check at the beginning of the trial failed for 3 consecutive times and after each 10 trials.

4.2.3 *Corpus dataset*

In our corpus dataset subjects were shown 90 images to 105 participants in three groups with slightly varying viewing conditions asking them to remember which images they had seen for a subsequent test.

Participants

For this study 105 students of the University of Potsdam with normal or corrected to normal vision were recruited. On average participants were 23.3 years old and 89 of the participants were female. Participants received credit points or a monetary compensation of 16 Euro for their participation. The work was carried out in accordance with

the Declaration of Helsinki. Informed consent was obtained for experimentation by all participants.

Stimuli

As stimuli we selected 90 photographs, which did not contain prominent humans or text. Furthermore images were selected as 6 subsets of 15 images each: The first contained photographs of texture-like patterns, the other 5 contained typical holiday photographs with the prominent structure either at the top, left, bottom, right or center. The full set of images is available online with the dataset.

For presentation in grayscale I measured the luminance output $[\frac{cd}{m^2}]$ of each gun separately and for the sum of all three guns at every value from 0 to 255. To convert a stimulus into grayscale I summed the luminance output for the RGB values and chose the gray value with the most similar luminance.

Procedure

Eye movements for our Corpus experiment were collected in two sessions. In each session 60 images were presented and participants were instructed to memorize them for a subsequent test to report which images they had seen. In the first session all images were new. In the second session 30 images were repeated from the first session and the remaining 30 new images were shown. The 30 repeated images were the same for each observer. For this Chapter I use all fixations from both sessions ignoring whether the subject had seen the image before and which group the subject belonged to, to maximise the amount of data. Trials began with a black fixation cross presented on a gray background. After successful binocular fixation in a square with a side length of 2.2° the stimulus appeared and subjects had 10 seconds to explore the image. In the memory test participants had to indicate for 120 images if they had seen them before. Half the images were the ones they saw in the experiment, the other half were chosen randomly from another pool of 90 images was chosen according to the same criteria as the images used for the first set of images.

The three cohorts of subjects differed in the placement of the fixation cross and whether the images were shown in colour or in grayscale:

- For the first 35 subjects images were presented in grayscale and the start position was placed randomly within a doughnut-shape around the center of the screen and stimulus with an inner radius of $100px = 2.6°$ and an outer radius of $300px = 7.8°$

- For the second group of 35 subjects the images were also presented in grayscale, but the start position was chosen randomly from only 5 positions: The image center and 20% of the monitor size (256/205 pixels, 5.68/4.55 degree of visual angle) away from the border of the monitor at the top, left, bottom and right centrally in the other dimension.

- For the final group of 35 subjects the images were shown in colour and the starting position was as for the second group.

4.2.4 *Natural image search*

Participants

Eye movements were recorded from 10 human participants (4 female) with normal or corrected-to-normal vision in 8 separate sessions on different days. 6 participants were students from a nearby high school (age 17 to 18) and 4 were students at the University of Potsdam (age 22 to 26).

Stimuli

As natural image backgrounds 25 images were taken by our collaborators in the area surrounding Potsdam, which contained no faces or writing.

As targets I designed 6 different low level targets with different orientation and spatial frequency content (Figure 28). To embed the targets into the natural images I first converted the image to luminance values based on a power function fitted to the measured luminance response of the monitor. Then I combined this luminance image I_L with the target T with a luminance amplitude αL_{max} fixed relative to the maximum luminance displayable on the monitor L_{max} as follows:

$$I_{fin} = \alpha L_{max} + (1 - 2\alpha) I_L + \alpha L_{max} T \tag{28}$$

, i.e. I rescaled the image to the range $[\alpha, (1-\alpha)]L_{max}$ and then added the target with a luminance amplitude of αL_{max}, such that the final image I_{fin} never left the displayable range.

Finally, I converted the image I_{fin} back to $[0, 255]$ grayscale values by inverting the fitted power function.

α was set to 0.15 by Lars Rothkegel to achieve roughly 80% correct in a pilot study on himself.

Procedure

Participants were instructed to search for one of 6 targets for the upcoming block of 25 images. To do so, the target was presented on a 26th demonstration image, marked by a red square. Each session consisted of 6 blocks of 25 images for each of the 6 different targets. The 25 images within a block were always the same presented in a new random order.

Trials began with a black fixation cross presented on gray background at a random position within the image borders. After successful fixation, the image was presented with the fixation cross still present for 125 ms. This was done to assure a prolonged first fixation to reduce the central fixation tendency of the initial saccadic response (Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2017; Tatler, 2007). After removal of the fixation cross, participants were allowed to search the image for the previously defined target for 10 s. Participants were instructed to press the space bar to abort the trial once a target was found. In $\approx 80\%$ of the trials the target was present.

At the end of each session participants could earn a bonus of up to 5€ additional to a fixed 10€ reimbursement, depending on the number of points collected divided by number of possible points. If participants correctly identified a target they earned 1 point. If participants pressed the bar although no target was present, one point was subtracted.

4.2.5 *Analysing fixation locations*

I extracted 79×79 pixel patches ($\approx 2.05 \times 2.05$ dva), around the fixation location, for all fixation locations for which this patch lay entirely inside the image. To obtain comparison patches, I extracted patches at the measured fixations locations shifting the image index by one. For example, I used the fixations from picture one to extract patches from picture two and so on and the fixations from the last picture to extract patches from the first picture. This method has been used in the past to train saliency models (e.g. Judd et al., 2009; Kienzle et al., 2009).

For analysis the patches were first converted to luminance. Following this conversion I computed spectra of the image patches and processed them with our early vision model using the contrast sensitivity function for 300 ms presentation time, the parameters I fitted for $79ms$ presentations times and mean normalization (See Chapter 2 for details on the model. For display I compute the average amplitude spectra and early vision responses per target and divide by the average over the control patches or over all targets.

4.2.6 *Gold standard analyses*

To estimate empirical fixation densities, I used kernel density estimation as implemented in the R package SpatStat (version 1.51-0).

To estimate the bandwidth for the kernel density estimate I use leave one subject out cross-validation, i.e. for each subject I evaluate the likelihood of their data under a kernel density estimate based on the data from all other subjects. For the image dependent density estimates I repeat this procedure with bandwidths ranging from .5 to 2 degrees of visual angle (dva) in steps of 0.1 dva. I report the results with the best bandwidth chosen for each image separately. For the image independent prediction—i.e. the central fixation bias—I use the same procedure with bandwidths from 0.2 to 2.2 dva as these estimates are based on more data and chose a single bandwidth over all images.

For my analysis over time I calculate two separate estimates of the gold standards. For the first I simply took the cross-validated kernel density estimate based only on the fixations from each ordinal fixation number (labelled "Empirical Saliency"). This first estimate declines rapidly over time, as fixations typically become fewer and more dispersed later in the trial. To counteract this I compute a second estimate which uses all fixations on the image from the second to the last to predict the density (labelled "Empirical Saliency All Fixations"). This estimate can use more data and performs well, because the fixation density converges towards the end of the trial (see Fig. 31).

4.2.7 *Comparing fixation densities*

To compare two fixation densities p_1, p_2 I compute a kernel density estimate $\hat{p}_1$ for one of the fixation densities p_1 and evaluated the log-likelihood of the fixations $f_2^{(i)}$ measured for the other fixation density. The following equations show that this is an estimate for the negative of the cross-entropy of the two densities $H(p_2; p_1)$.

$$H(p_2; p_1) = -\int p_2(x) \log(p_1(x)) dx \tag{29}$$

$$= -E_{p_2}(\log(p_1(x))) \tag{30}$$

$$\approx -\frac{1}{n} \sum_{i=1}^{n} \log\left(\hat{p}_1(f_2^{(i)})\right) \tag{31}$$

The cross-entropy is closely related to the Kullback–Leibler divergence $KL(p_2||p_1)$, which is simply the cross-entropy minus the entropy of p_2, i.e. in a formula:

$$KL(p_2||p_1) = H(p_2; p_1) - H(p_2) \tag{32}$$

Thus the log-likelihood I report measures how well p_1 approximates p_2, irrespective of the entropy of p_2, i.e. irrespective of the upper limit for predictions of p_2.

COMPARISONS OVER TIME I use this log-likelihood measure to compare fixation densities over time taking the distributions of fixations with two given ordinal fixation numbers as p_1 and p_2. I again tried different bandwidths for the kernel density estimate and report the maximal value after leave one subject out cross-validation averaged over images with the optimal bandwidth chosen for each image.

COMPARISONS BETWEEN TARGETS For the search data I additionally compare the fixation densities produced by subjects when searching for different targets on the same images. As for all comparisons, the best performing bandwidth after leave one subject out cross-validation was chosen for each image and we report the average over all subjects and images.

4.2.8 *Evaluation of saliency models*

In our analysis of saliency models I largely follow Kümmerer et al. (2015), who recommend to use the log-likelihood of fixations under the model for evaluation after fitting a non-linearity, blur and center bias for each model to map the saliency map to an optimal prediction for the fixation density.

To fit this mapping from the saliency map to the fixation density I used the deep neural network framework Keras as included in tensorflow (Abadi et al., 2015, v. 1.3.0).

In this framework I fit a shallow network as illustrated in Figure 29 for each saliency model separately after resizing the saliency maps to 128×128 pixel resolution and rescaling the saliency values to the interval $[0, 1]$.

The network contained two conventional 1×1 convolution layers which first map the original to an intermediate layer with 5 channels and then to a single output layer, allowing for a broad range of strictly local non-linear mappings to the fixation density.

Next I applied a blurring filter and the activations were passed through a sigmoidal non-linearity to map them to strictly positive numbers. For the blur I implemented a 25×25 custom convolution layer, in which I set the weights to a Gaussian shape of which I fitted the two standard deviations.

Finally I apply a center bias through a custom layer. This layer first multiplies the map with a Gaussian with separately fitted vertical and horizontal standard deviations

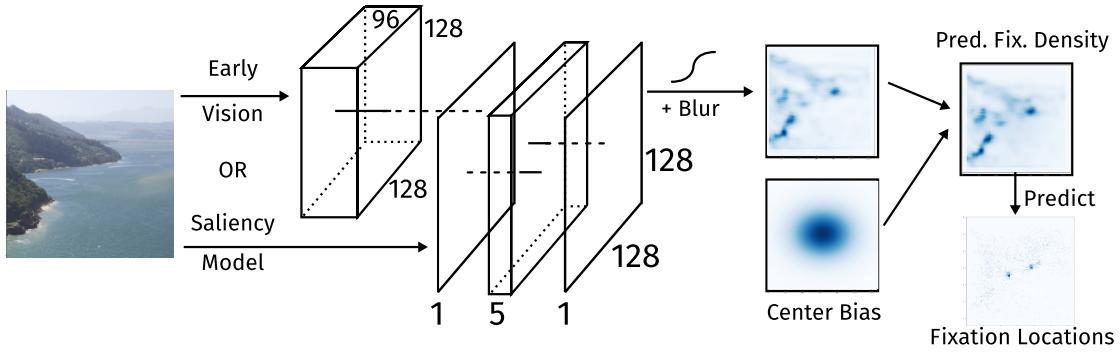

Figure 29: Shallow neural network to map raw saliency models to fixation densities. I first compute a raw saliency map from the image, either by applying the saliency model or by linearly weighing the 96 response maps produced by our early vision model. Then two 1×1 convolutions are applied which first map the values to 5 intermediate values per pixel locally and then map to a single layer with a Relu non-linearity inbetween, which effectively allows a piecewise linear map with 5 steps as an adjustable local non-linearity. I then apply a fixed sigmoidal non-linearity and blur with a Gaussian with adjustable size. Finally I multiply with a fitted Gaussian Center Bias, which results in the predicted fixation density, which can be evaluated based on the measured fixation locations.

and then normalizes the sum of the activities over the image to 1 to obtain a probability density.[1]

As a loss I directly use the log-likelihood as for the kernel density estimates described above. In Keras I implemented this by flattening the final density estimate and using the standard loss function *categorical_crossentropy* to compare to a map with sum 1 and entries proportional to the number of fixations at each pixel location.

For evaluation I performed 5-fold cross-validation over the used images, i.e. I trained the network 5 independent times leaving out one fifth of the data. For training I used the Adam optimization algorithm (Kingma & Ba, 2014) with standard parameters till convergence by reducing the learning rate by a factor of 2 whenever the loss improved less than 10^{-5} over 100 epochs and stopping the optimization when the loss improved by less than 10^{-6} over 500 epochs. I did not employ a test set here as I did not optimize any hyper parameters and did not use any stopping or optimization rules based on the validation set.

4.2.9 *Tested saliency models*

To get a comprehensive overview over saliency model performance, I chose a few representative models for predicting saliency:

KIENZLE As an example of a extremely low level model of visual saliency I employ the model by Kienzle et al. (2009), using the original implementation supplied by Felix Wichmann.

ITTI & KOCH As the most classic saliency model I evaluated the original model by Itti et al. (1998). To compute the saliency maps I used the implementation which ac-

1 I also implemented an additive center bias, which performed worse than the multiplicative version for all models, however.

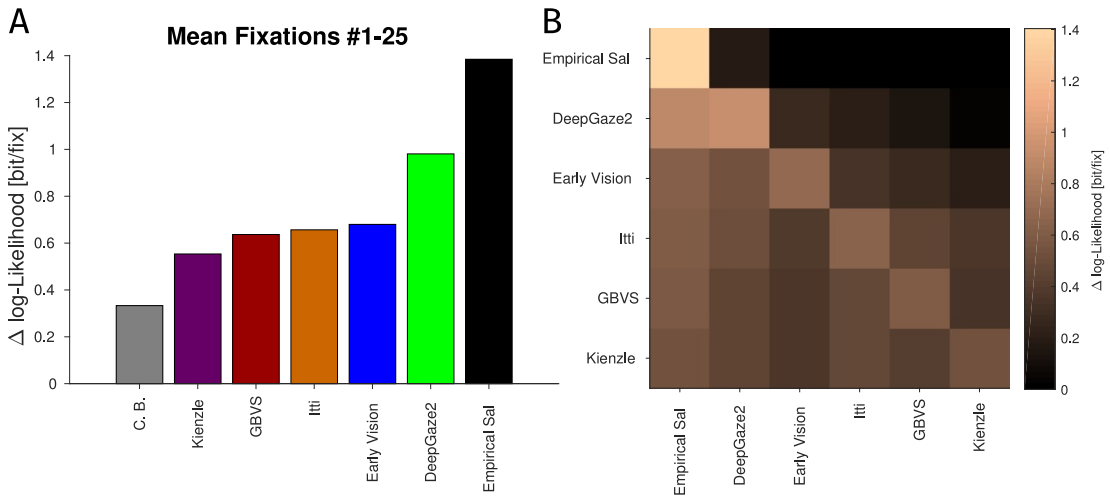

Figure 30: A: Average performance of the models. B:Similarity of the different saliency maps. Measured in terms of Δ log-likelihood, i.e. as the prediction quality when using one map to predict random draws from another.

companies the GBVS saliency model, which performed decisively better than the original implementation from www.saliencytoolbox.net.

GBVS As a better performing classical hand crafted saliency model I use the Graph based visual saliency model by Harel et al. (2006). Code was downloaded from here[2].

DEEPGAZE II As a representative of the newest deep neural network based saliency models I chose DeepGaze II by Kümmerer et al. (2016). This model is currently leading the MIT-saliency benchmark (Bylinskii et al., 2016). Saliency maps for this model were obtained from the webservice at deepgaze.bethgelab.org as log-values in a .mat file and converted to linear scale before use.

EARLY VISION Our early vision saliency model is based on our psychophysical spatial vision model I published recently (Schütt & Wichmann, 2017). This model implements the standard model of early visual processing to make predictions for arbitrary luminance images. As an output it produces a set of $8 \times 12 = 96$ orientation $\times$ spatial frequency channel responses, spatially resolved over the image.

To obtain a saliency map from these channel responses I linearly weighted and added them to form a saliency map which I plugged into the same machinery as the saliency maps for all other models. This implementation allows us to train an arbitrary weighting for the maps from our early vision model directly while keeping the benefits of a non-linearity, blur and center bias as for the other models.

4.3 RESULTS: CORPUS

4.3.1 *Overall saliency model performance*

To test our models for visual saliency based on a model of spatial vision I evaluated their performance along with a range of classical low-level saliency models (Itti& Koch, Itti et

2 Link for the paper version: http://www.vision.caltech.edu/ harel/share/gbvs.php

al. (1998), GBVS, Harel et al. (2006), Kienzle, Kienzle et al. (2009)) and to the currently best DNN-based Saliency model DeepGaze II (Kümmerer et al., 2016). For these models we fitted the same non-linear map, blur and center bias as for our early spatial vision based model to make them comparable, as described in methods. As the evaluation criterion we use the average log-likelihood difference to a uniform model as described by Kümmerer et al. (2015) for saliency models and Schütt et al. (2017) for dynamical models.

The results of the overall analysis [3] are displayed in Figure 30. My early vision based saliency model performs slightly better than the classical bottom-up saliency models using only a simple weighted sum of activities as a saliency map. Thus such a simple sum seems to be sufficient for modelling bottom-up low-level influences.

However, DeepGaze II ist clearly the best model, with a substantial $0.3\frac{bit}{fix}$ better than all classical saliency models & our early vision based models. However, DeepGaze II is not as close to perfect prediction for our corpus dataset as for the MIT saliency benchmark, missing it by roughly $0.4\frac{bit}{fix}$ (compare our Figure 30 B to Kümmerer et al. (2016), Figure 3). A potential reason for this might be that our dataset contains many more fixations per image (≈ 2600), than the saliency benchmark (Judd, Durand, & Torralba, 2012, 39 observers $\times$ 3 seconds ≤ 390), which allows a more detailed estimation of the empirical fixation density and thus a higher gold standard. An alternative explanation is that the mit-benchmark dataset contains (more) humans, faces and text, which might help DeepGaze II, as these are typical high-level properties reported to attract fixations.

These overall performance results suggest that a realistic early vision representation indeed provides similar predictive value for the density of fixations as classical saliency models do, although good fixation predictions require more. This does not fully answer the question whether classical saliency truly represents early visual processing though. To approach this question, I additionally analysed how similar the predictions of the different saliency models are. To compare saliency model predictions I calculated the cross-entropies between the different predicted fixation densities on the same scale I used to evaluate the performance of the models before.

The resulting cross-entropies between saliency models are shown in Figure 30 B. The empirical saliency predicts itself more accurately than any saliency model predicts itself, i.e. it has the lowest entropy. Also, each of the saliency models is distinct from the others, as the diagonal elements are larger than any corresponding off-diagonal ones. However there is a group of models which make similar predictions: The classical saliency models except share some common entropy. These results imply that the early vision model based saliency is somewhat different from the classical saliency models. Finally, we can observe some asymmetries in the prediction qualities. For example, the empirical fixation density is predicted reasonably by the saliency models, but does not work well as a prediction for the saliency maps.

This pattern confirms that the fixation density is more concentrated than the predictions and all saliency models predict fixations in areas which are in truth rarely fixated. A similar, but weaker relationship also exists between DeepGaze II and the other saliency models, as DeepGaze II is better predicted by the other models than the other way around. The tendency that more successful saliency models generally become more specific than less successful ones is partially caused by the link to fixation density I fit. Bad predictions are weighted weaker than good ones, such that the density automatically becomes broader.

3 I only plot fixations #1-#25 here for consistency with later plots over time. Including later fixations does not qualitatively change the results displayed here. The only change is that later fixations are predicted worse by all models decreasing the absolute performance of all models slightly.

4.3.2 *Temporal aspects*

To evaluate the saliency models over time I split our dataset by fixation number within a trial. I then computed a kernel density estimate for each fixation number and evaluated the likelihood of the fixations of each fixation number separately. For these estimates I used a bandwidth of 1.6°, because it gave the highest likelihood for the average over fixations 2-25. The results of this analysis averaged over images are displayed in Figure 31 A. Caused by the cross-validation over subjects, the estimates for each fixation number predicting itself are interpretable and comparable to the ones where other fixation numbers are predicted.

Going through the plot in temporal order I find that:

1. The 0th fixation (the start position) neither predicts the other fixation locations nor is predicted by them well.

2. The first and to a lesser degree the following fixations show an asymmetric pattern: They predict other fixations badly, but are predicted well by other fixation numbers, indicating that they land at positions which are fixated later as well, but do not cover all of them.

3. This tendency gradually declines from the second fixation till roughly the 10th fixation accompanied by a gradual decline in predictability.

4. From the 10th fixation onwards the fixation densities of all fixation numbers predict each other equally well, indicating that the fixation density has reached an equilibrium state.

These results suggest a separation into three phases: The first fixation, which seems to be different from all others, the phase with the asymmetric pattern when fixations are well predicted by the later density, have not converged to it yet and the final equilibrium phase when the fixation density has converged.

Our next aim was to quantify how good predictions based on an image could possibly be at different time points after image onset. To quantify this, I used four limiting cases: First, a central fixation bias implemented as a kernel density estimate from fixations from all images with the correct fixation number. Second, a central fixation bias based on all fixations from all images. Third, the empirical saliency estimated as a kernel density estimate from the fixations with the same fixation number on the same image. Fourth, a different estimate of the empirical saliency estimated from fixations number 2 to 25 on the given image, to increase the number of fixations available for the kernel density estimation.

The results of this analysis are displayed in Figure 31 B. The image independent prediction declines quickly from the good prediction based on the initial central fixation bias on the first fixation to the constant level of roughly $0.25 \frac{bit}{fix}$, which is retained over the whole trial. Also the two estimates only differ for substantially for the first few fixations affected by the initial central fixation bias. For the empirical saliency, both estimates show a gradual decline over time. The estimate based on all fixation numbers flattens out between the 10th and 15th fixation while the one based only on the fixations with the same fixation number keeps decreasing, most likely due to the decreasing number of available fixations. However, this fixation specific empirical density reaches a much higher value for the first fixation, reiterating that the first fixation follows a different density than later ones.

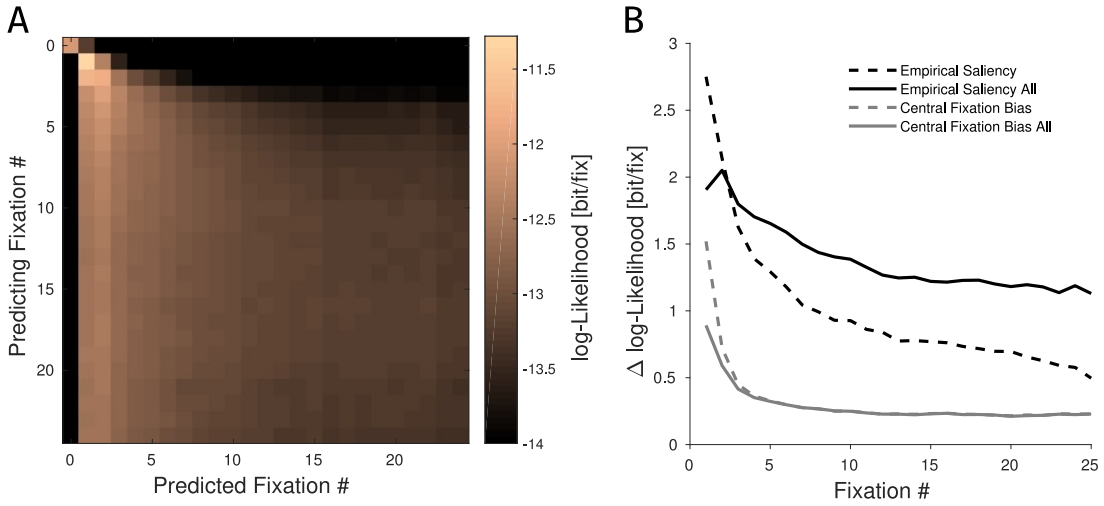

Figure 31: Analysis of the predictability of fixation densities over time. A: log-likelihood for predicting the fixations with one fixation number from fixations with a different fixation number as a measure how well the density at one fixation number predicts the fixations with a second fixation number. B: Performance of the Gold Standards over time. Shown are the performance of the empirical saliency measured by predicting the fixations of one subject from the fixations of other subjects and the central fixation bias measuredd by predicting the fixations in one image based on the fixations in other images. For each of these limits two curves are shown: One continuous line based on only fixations with this fixation number and one dashed line based on all fixation numbers.

I interpret this observation as further evidence for a separation into a short initial central fixation bias dominated period, a period for which predictability gradually declines and a late equilibrium period. Additionally, the difference between our two estimates of the maximally predictable information shows that the ~ 100 fixations we have for each fixation number are not enough for a good estimate of the fixation density. Thus the fixation density estimate from all later fixations gives a better estimate of the maximally attainable fixation density for all but the first and possibly second fixations which seem to deviate from what attracts later fixations.

4.3.3 *Saliency models over time*

The performance of saliency models over time is of interest to test the prediction that low level features play a more important role at the beginning of a trial. The results of this evaluation are displayed in Figure 32. In general, the prediction quality of the saliency models follows the curve for the Empirical saliency with a gradual decline which reaches a plateau between the 10th and 15th fixation. As expected, all saliency models are better than a central fixation bias but do not explain the fixation density perfectly yet.

The differences between the different models I observed in their overall performance are present throughout the trial. DeepGaze II performs best and the other saliency models run largely in parallel, $\approx 0.4 \frac{bit}{fix}$ below. To investigate the additional contribution of high-level features, I plot the difference between DeepGaze II and the early vision based model in Figure 32 B. This plot emphasises that DeepGaze II is always predicting fixations better than the early vision based model, although the first fixation shows a somewhat smaller advantage of DeepGaze II. Already at the second fixation the difference between DeepGaze II and our model is largest and then roughly follows the decline in general

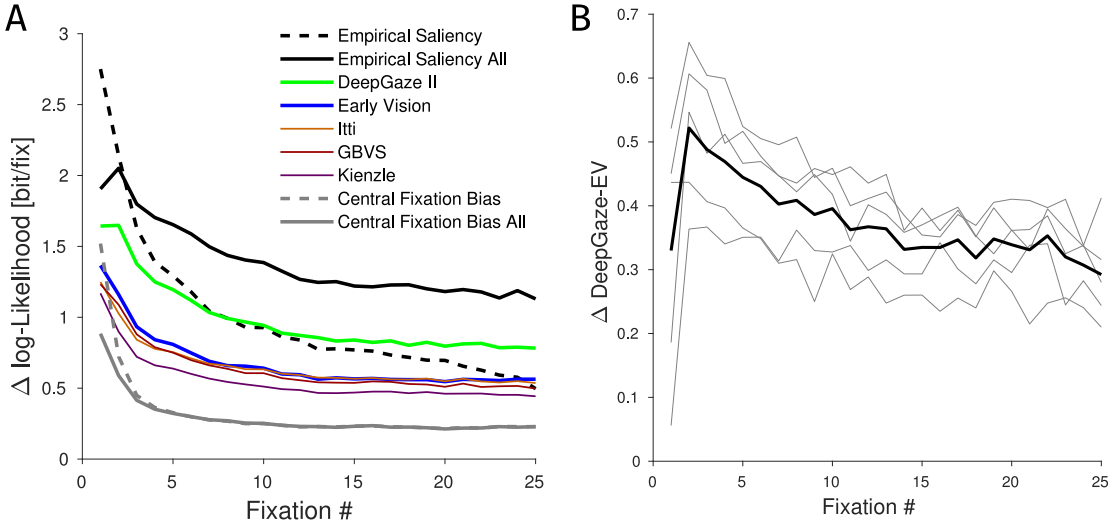

Figure 32: Saliency Model Performance on the Corpus. A: Performance of the saliency models over time, replotting the maximal achievable values from Figure 31. B: Difference between DeepGaze II and the early vision model over time. The gray lines represent the individual folds.

predictability of fixation locations until it converges somewhere between fixation # 10 and # 15.

Thus the general observation that saliency models based only on low level features perform worse than models which include object information can be confirmed throughout the trial. The only fixation for which the advantage of object based information might be smaller is the first one. As this fixation also contains a strong central fixation bias, which varies over time (Rothkegel et al., 2017) and was proposed as the main point in time for saliency effects (Anderson et al., 2015), I analyse this first fixation in more detail below.

4.3.4 *Density of the first fixation*

To analyse the first fixation in detail I performed two complementary analyses: I display some raw data in Figure 33. And I compare the performance of our early vision based model and DeepGaze II to the performance of the center bias and the empirical density prediction over time within the first fixation in Figure 34. For each predictor I fit the first fixation and all other fixations separately. For the saliency models I retrained our network, i.e. learned a separate blur, non-linearity and center bias. For the empirical density and center bias I generated separate kernel density estimates.

Generally the density for the first fixation shows a pronounced *initial* center bias (Rothkegel et al., 2017; Tatler, 2007), i.e. early saccades almost exclusively move towards the center of an image. This tendency is visible in the raw data (for example in the upper left image in Figure 33) and also in the much better prediction quality of the image independent central fixation bias model (see Figure 34).

Nonetheless, I observe that the first fixation is clearly guided by the image. I find that fixations can be predicted much better when including knowledge about the image (see Figure 31 & 34) and can confirm this by looking at examples in Figure 33. I can also confirm the observation that the first fixation differs from later fixations as all predictions fitted to the first fixation separately perform much better than the fits to the later fixations.

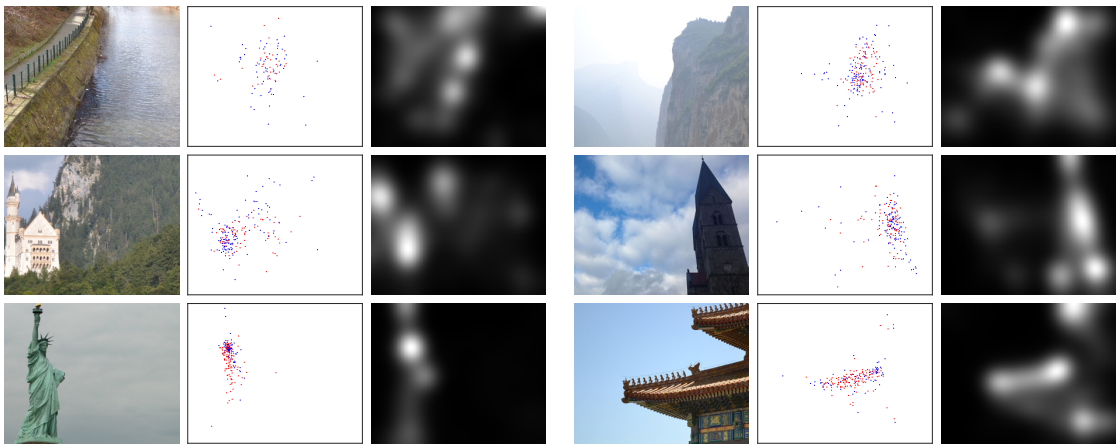

Figure 33: Examples showing the differences among images in the initial central fixation bias. For each image I show the image, the first chosen fixations as a scatter plot and the density of all later fixations. Colour represents a median split by the fixation duration at the start location, red fixations were chosen after less than $270ms$, blue fixations after more than $270ms$. The left column shows examples of our left focussed images, the right column of the right focussed ones.

This observation can be confirmed referring to the raw data as well, as the first fixations do not seem to follow the density I computed from later fixations (see Figure 33)

Also, DeepGaze II performs better than the early vision based model although the difference is a bit smaller than at later fixations and all models perform much better, such that the difference is relatively smaller.

Analysing the temporal evolution, all predictions are relatively bad for first fixations with latencies below $150ms$, which appear not to be guided by the image yet, but represent only 5% of first fixations. After this bad performance follows the bulk of fixations between 200 and 400 ms which are best predicted by all models. These fixations already show an advantage of the DeepGaze II model. After this period prediction quality declines for the models trained for the first fixation emphasizing that late saccades follow a different density than earlier ones. The models trained on the later fixations decline much slower. This slower decline for the late trained models could be the earliest part of the general decline in predictability we observe over multiple fixations above. Thus, later first saccades might already follow the same factors as later fixations.

Interpreting these results I conclude that high-level information is advantageous for the prediction of eye movements already 200ms after image onset. However, the central fixation bias and the low-level guidance are much better models for the first fixation than for later ones and the advantage of using high-level information is smaller for the first fixation. Thus the first fixation is still most likely to contain bottom-up low-level guidance.

The simplest models to explain the central fixation bias would be to add a certain proportion of fixations driven by the initial central fixation bias or to reweigh the fixation density depending on the distance to the center. However, exploring the first chosen fixations in more detail shows at least two problems with these accounts, illustrated by the examples shown in Figure 33. First, the strength of the central fixation bias differs considerably between images. For some images the fixations are indeed consistent with a Gaussian distribution around the image center (e.g. top left). For others the image content dominates even the choice of the first fixation (e.g. bottom left). Second, where

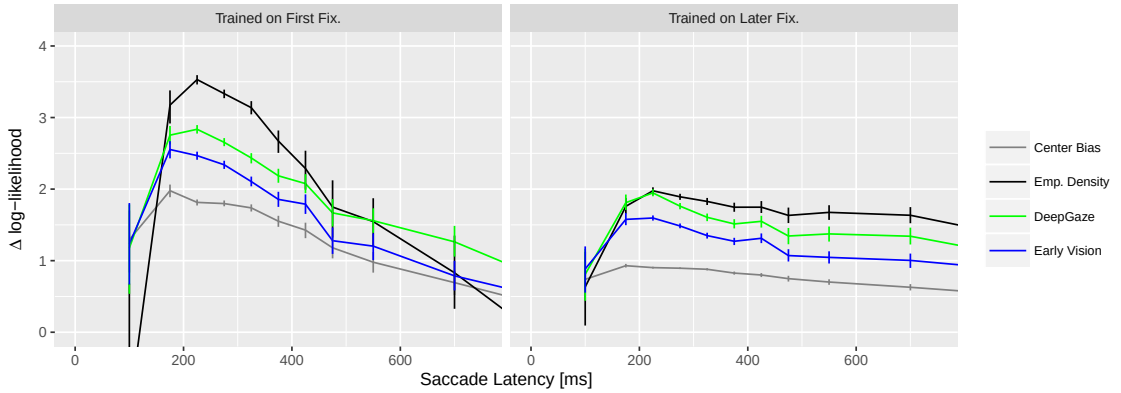

Figure 34: Temporal evolution of prediction qualities during the first fixation. I plot the log-likelihood gain compared to a uniform distribution for empirical density, center bias, early vision based saliency model and DeepGaze II. For display saccade latencies were binned, errorbars represent bootstrapped 95% confidence intervals for the mean.

the first fixations depend on the image, this distribution still differs from the later fixation distribution in some images, at least in the weighting of the different targets (e.g. right middle and bottom). Thus both reweighing and added fixations appear to be necessary and different images need to receive a differently strong central fixation bias.

4.4 RESULTS: VISUAL SEARCH

4.4.1 *Fixation densities*

The first questions I asked about the visual search data was whether fixation locations are predictable from the image and how different the fixation densities are for the different search targets. To investigate this, I calculated kernel density estimates from the fixation locations for each search target. Then I used this estimate to evaluate the cross-validated likelihoods of the data for the same target or any of the other targets. This calculation estimates the negative (cross-) entropies of the fixation locations relative to a uniform distribution.

The results are displayed in Figure 35. In panel A the gold standard and center bias performance is shown for the different targets. Comparing these likelihoods to the free viewing data it is clear that the fixations chosen during search are distributed much broader over the images than the ones chosen during free viewing. The individual targets each reach $0.5 - 0.6\frac{bit}{fix}$ of predictable information and the shared information is only $0.3 - 0.4\frac{bit}{fix}$, while the empirical fixation density explained $\approx 1.4\frac{bit}{fix}$ for the free viewing data. In panel B the log-likelihood relative to that of a uniform distribution is displayed. The fixation densities for the targets separate into three groups. The three high spatial frequency targets lead to similar fixation distributions. Furthermore, the Gaussian blob and the positive Mexican hat lead to similar distributions, while the negative Mexican hat produces a different distribution from all others. Nonetheless, the log-likelihood of the fixations for any target were higher under the fixation densities estimated for any other target than for the uniform distribution (all cells $\gg 0$). This indicates that some areas attract fixations independent of the target subjects search for.

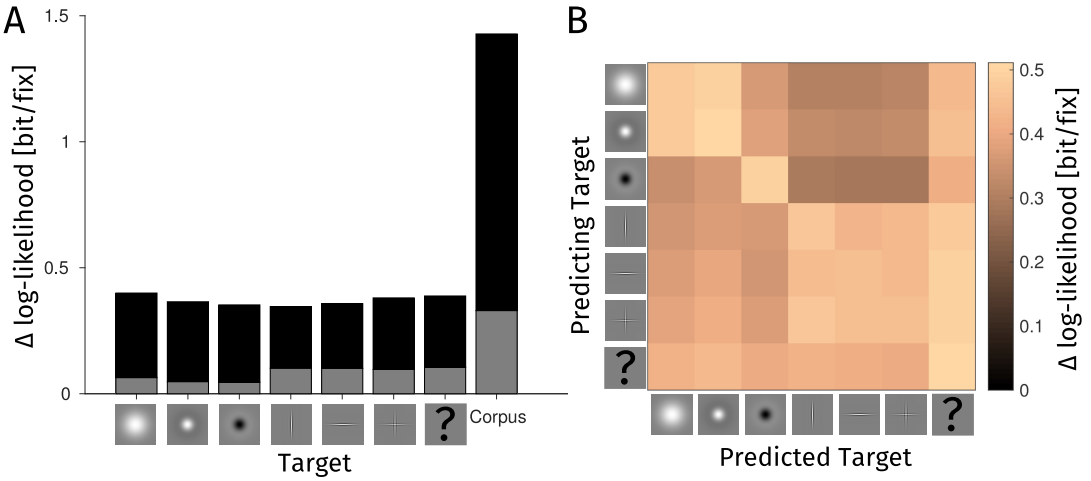

Figure 35: Analysis of fixation densities in the search experiment. A: Prediction limits for the fixation densities for the different search targets estimated from leave one subject out cross-validation. The gray lower proportion indicates the maximum for image independent prediction (central fixation bias). The black bars represent the maximum for image (& target) dependent prediction. I additionally plot these values for the Corpus dataset for comparison. B: Δ log-likelihood as a measure of prediction quality when predicting the fixation locations when searching for one target from the fixation locations when searching for a different target in the same image.

Thus, there is something to be predicted, although it is much less than in the free viewing experiment and observers change the locations they fixate depending on the target they search for, corroborating our earlier observation that searchers adjust their eye movement dynamics to the target they search for (Rothkegel et al., 2018).

4.4.2 *Saliency models*

For the analysis of the saliency models I employed the same techniques as for the Corpus dataset. I fit a non-linearity, blur and central fixation bias and evaluate the performance of the resulting prediction over time using cross-validation.

As shown in Figure 36, no saliency models predicts the fixation density well during visual search beyond the first few fixations. When the density prediction are not adjusted to the search data the models are worse than a uniform prediction at most timepoints. The only time these densities have any predictive value are the first and second fixations when there is an initial central fixation bias. When the connection from saliency map to fixation density was trained newly for the search data, the saliency models still explain only a tiny fraction of the fixation density. Even DeepGaze II and the version of (Itti & Koch, 2001) provided with GBVS, which perform best explain less than 0.2 $\frac{bit}{fix}$, i.e. less than a third of the explainable information. Adjusting the link even stronger, I also trained the connection from saliency to fixation density separately for each target. This had little effect for any of the saliency models and the early vision based model did not profit from this adjustment either although its performance changed slightly and indeed improved on the training dataset at least (not shown).

Furthermore I evaluated the DeepGaze II model—which performed best for free viewing—without the link I provided (shown as 'DeepGaze2 raw'). This evaluation is possible, because this model—contrary to the other models—already predicts a density as its saliency

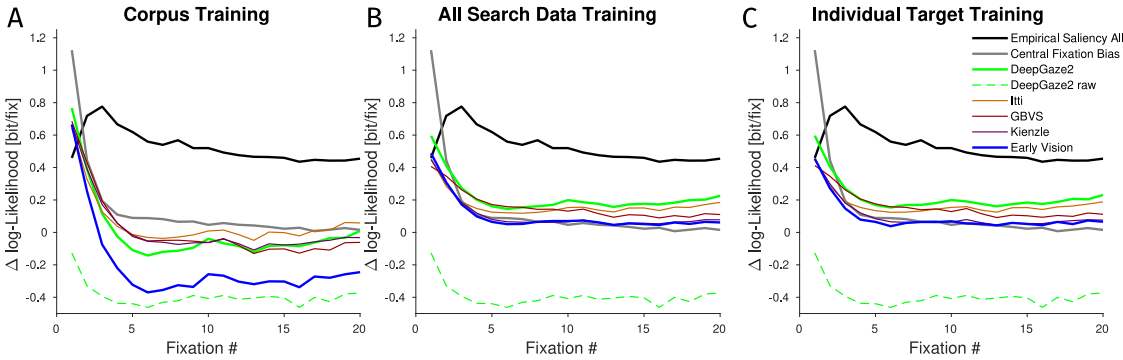

Figure 36: Performance of the Saliency models on the Search dataset over time. The different columns show different conditions for training the connection from saliency map to fixation density. Free-viewing training: taking the mapping which I trained for the free viewing corpus. All search data training: Using all Search data from the training folds. Inidividual target training: Training and evaluation was performed separately for each search target; Results are averaged over targets. Additional to the different saliency maps I also plot the empirical saliency performance, the center bias performance fitted per fixation number and the performance of the unmodified DeepGaze II saliency map (DeepGaze2 raw).

map. The raw prediction of DeepGaze II is clearly below chance performance as well, emphasising that the link I fitted here is not responsible for the failure of this model.

Thus our results confirm that the fixation locations during visual search are neither predicted by any bottom-up model nor by low-level features, whether they are adjusted to the task or not.

4.4.3 *Fixated patches*

Earlier analyses of eye movements during visual search reported similarities between the fixated locations and the target and it was usually assumed that such relationships are exploitable for the prediction of fixation locations. Given the failure I observed for predicting eye movements, I wanted to check whether the corresponding differences between fixated and non-fixated image locations exist in the dataset. As fixated locations, I extracted patches around the fixation locations and compared them to control patches extracted from the same locations in a different image from the stimulus set (see Methods), as often been done before (e.g. Judd et al., 2009; Kienzle et al., 2009).

As displayed in Figure 37A, the average spectrum of a fixated patch looks much like the spectrum of any image patch with a clear $1/f$ decline in spatial frequency content and a preference for horizontal and vertical structure. As these strong effects hide all other effects, all other spectra are divided by the spectra of the comparison patches for display.

The overall spectrum of fixated patches shows increased power for all frequencies and orientations (Fig. 37 B). Searching for a specific target additionally produces a slight bias of the fixated image patches towards being more similar to the spectrum of the target (Fig.37 D). The deviations of the single targets from the grand average are all smaller than 5%, however, while the variance over patches is substantial ($\frac{SD}{M} \in [78.65\%, 161.03\%]$, average $= 91.10\%$). The unknown target condition (Fig. 37 C) produces no clear deviation from the average over the conditions with known target.

These results confirm that our dataset is not simply an outlier lacking low-level guidance.

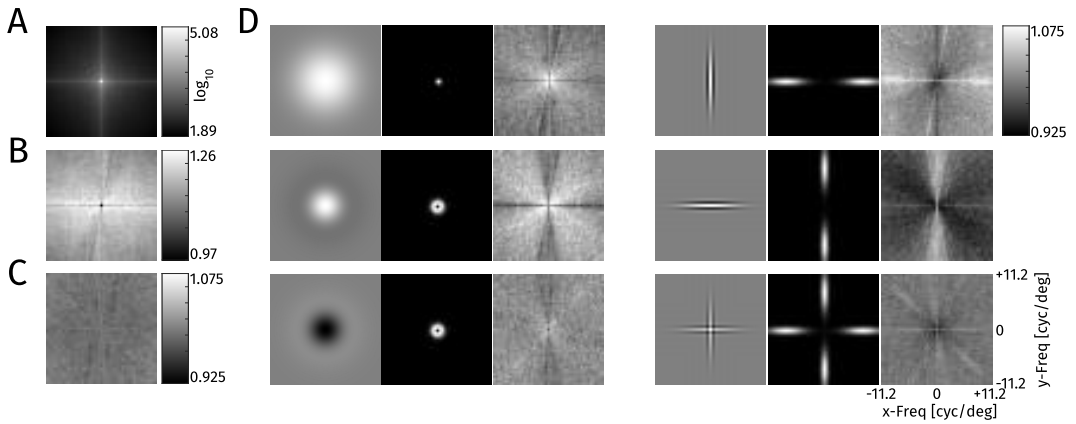

Figure 37: Analysis of the spectra at fixated locations. **A**: Grand average spectrum over all fixated patches. **B**: A divided by the average spectrum at control locations. The value at 0 frequency is 0.97, all other values are in the range $[1.09, 1.26]$ **C**: Average spectrum for fixations when the target is unknown, plotted as for known targets in D. **D**: Triples for each target: The target at 100% contrast against a gray background, the Fourier space representation of the target and the average spectrum divided by the average over all targets. The color range from black to white for the third plot is always $[0.925, 1.075]$.

4.4.4 *Predicting search performance*

As the prediction of fixation locations based on the early vision model responses failed, I wanted to check that my model is at least predictive of search performance in the sense that targets which the model predicts to be easier to discriminate from the background are actually found more often and faster in our search task. This result is predicted by any account of visual search in natural scenes. Thus, this connection may serve as a validation of my early spatial vision model, whose predictions were checked only in much less natural discrimination tasks in Chapter 2.

As results of this analysis I plot the search performance for each specific target at a specific location against the signal to noise ratio (SNR) for detection predicted by our early spatial vision model in Figure 38. As measures for search performance I use the proportion of trials in which the target was found at that location and the average time used to find the target.

As expected search performance is predicted reasonably well by our early spatial vision model. Both measures of search performance are clearly correlated with the SNR predicted by the early spatial vision model for all targets.

Additionally the different targets lead to substantially different SNRs on average, while they were similarly hard to find in the experiment. Low spatial frequency targets are generally assigned lower SNR than the higher spatial frequency targets (compare the y-axes). I attribute this difference to the steeper decline in visibility for higher spatial frequencies. If peripheral information is important for visual search (as claimed by Rosenholtz, Huang, Raj, Balas, & Ilie, 2012, for example) this explains this difference, as the early vision model only covers foveal processing.

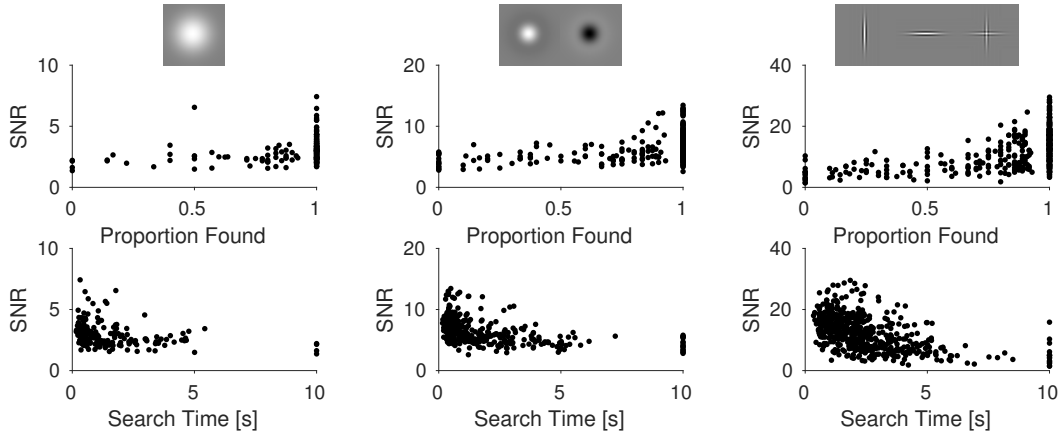

Figure 38: Predicting search performance based on visibility according to our early vision model. First row: Predicted signal to noise ratio (SNR) for detecting the target at the given location plotted against the proportion of trials the target was found. Second row: SNR against the average time required to find the target plotting targets-location combinations which were never found at the maximum duration of a trial of 10s. I split the targets in three groups (columns) here based on their spatial frequency content. The three groups show markedly different detectability according to our model (notice the change in y axis scale).

4.5 DISCUSSION

In this Chapter I explored the temporal evolution of the fixation density over the course of a trial. A more detailed look at the temporal dynamics of the fixation density can provide some insights how eye movements are controlled in a relatively natural environment. This analysis is made possible here for the first time by the long duration of trials and the large number of subjects for each image.

Based on the similarities of fixation densities shown in Figure 31 I propose a separation of a typical free viewing trial into three phases:

1. An onset response which affects mostly the first saccade.

2. The main exploration, which is characterized by a gradual broadening of the fixation density.

3. A final equilibrium state, in which the fixation density has converged.

I interpret these three phases as an initial orienting response towards the image center, which might be biased by large salient objects in the image, followed by a brief guided exploration during which observers look at all parts of the image they are interested in which is then followed by an idle phase during which observers look around rather aimlessly.

Exploring the onset response in more detail, I found some guidance beyond a simple movement to the image center. An image dependent prediction already performs substantially better than an image independent one (Figure 30). In fact, the first saccade is most consistent over subjects. Looking at examples of the fixation densities for the first saccade (Figure 33), the reader may confirm that these fixations are guided by the scene sensibly, despite fixating nearer to the centre. However, observers fixate central areas with their first saccade, which they do not fixate later in the trial. This is made visible in the bad prediction performance of the later fixation densities for the first one in Figure 31.

The main exploration focusses on similar image locations as the subjects fixate when the fixation density is converged (Figure 31). The fixations during this phase are even better predicted by the later fixation density than the later fixations themselves. During this phase the fixation density gradually broadens becoming less and less predictable. Correspondingly the performance of all saliency models is maximal at the beginning of this phase and decreases over time. Importantly DeepGaze II a model which includes high-level features, has the largest advantage at the beginning of this phase, i.e. the advantage of including high-level features starts immediately, at most $200ms$ after image onset.

Finally, in the last phase the fixation density reaches an equilibrium and all fixation numbers predict each other equally well. Although subjects preferentially return to the same fixation locations they visited during the main exploration they are overall less predictable.

In the search data I find a qualitatively similar temporal evolution of the fixation density as for memorization. Trials consist of an onset response with initial central fixation bias, a period of marginally better predictability and a final equilibrium state. However, the fixation density is much less predictable in general, the image independent prediction becomes entirely uniform and all saliency models perform much worse in predicting fixation locations, especially when the mapping from saliency map to fixation density is reused from the corpus dataset. Also the initial central fixation bias is weaker, which was expected as there was a delay of the first saccade in this dataset (Rothkegel et al., 2017). As the same models with the same mapping to the fixation distribution performed well on the free viewing data the failures on search data cannot be attributed to our method of linking saliency and the fixation density easily.

4.5.1 *Bottom-up vs. top-down*

Based on the search results I can confirm earlier reports that fixation locations during visual search are hardly predicted by saliency models (Chen & Zelinsky, 2006; Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007) which shows that top-down control can overwrite bottom-up control when subjects view static natural scenes. I even see some influence of the target subjects search for in our data, which argues for a fairly detailed adjustment of the eye movements to the concrete task at hand. This result fits well with earlier observations we made on this dataset (Rothkegel et al., 2018), which showed that subjects adjusted their saccade lengths and fixation durations to target they searched for as well. Thus our observations overall argue for a strong, detailed top-down influence on eye movement control.

This explanation implies that bottom-up factors usually have little effect driving eye movements. The only exception to this argument might be the first fixation chosen by the observer. The first chosen fixation follows a different density than later fixations. Also the saliency models perform best for the first fixations and even predict fixations in the visual search condition. The strongest bottom-up influence on the first fixation seems to be the central fixation bias. Nonetheless the advantage of saliency models over the central fixation bias is also largest for early first saccades $200 - 300ms$ after image onset.

Complicating the analysis of the first chosen fixation, I observe a temporal evolution within the first fixation, as we and others have observed before. Earlier saccades are more strongly biased towards the image centre (Rothkegel et al., 2017) and might be driven more by bottom-up features (Anderson, Donk, & Meeter, 2016; Anderson et al., 2015). This transition from bottom-up effects to more value driven saccades within a single

fixation duration was also observed in single saccade tasks with artificial stimuli (Schütz et al., 2012). Thus the transition from bottom-up to top-down control might occur very early, within the first fixation.

4.5.2 *Low-level vs. high-level*

At first glance, the observation that the low-level models predict fixations well at the beginning and worse later in the trial fits well with the classical saliency model idea that the initial exploration is driven by low-level bottom-up factors. However, the performance decline of low-level models follows the decline of the gold standard, early fixations are well predicted by the later fixation densities and the advantage of the DeepGaze II model is larger during the gradual broadening of the fixation density. These findings rather suggest that even during the initial exploration fixations are driven by the same high-level information later fixations are driven by. As all predictions decline in parallel, the reason for the decline might be an increase of fixations which are not guided by the scene at all.

Indeed, even within the first fixation adding high-level information improves predictions. Within $200ms$ after image onset DeepGaze II performs better than the early vision based model. While low-level factors may predict early first fixations better as has been observed before (Anderson et al., 2016, 2015), high-level factors seem to play a role from the start. Thus it appears that the influence of low-level features decreases for later fixations rather than the high-level influence increases.

This account agrees well with a range of literature which shows influences of objects (Einhäuser, Spain, & Perona, 2008; Stoll et al., 2015) and other high-level features (Henderson et al., 1999; Torralba et al., 2006) on eye movements. The predictive value of low level features like contrast at a location could then be explained by their correlation with being interesting in a high level sense. Such correlations are obviously existent in the sense that very low contrast areas are boring, because there is nothing to be seen. As such this explanation would also work to explain high level influences based on low level features. However, high level features are better at predicting, such that they necessarily have some predictive value beyond low-level features. Also manipulations of contrast seem to have little influence on the fixation distribution beyond the first fixation (Anderson et al., 2015; Açık et al., 2009), such that the part of the fixation distribution which could be explained both by low-level and by high-level features is more likely to be explained by high-level features.

In visual search, when eye movements appear to be under largely under top-down control fixations also seem to be based largely on high-level features: Our early vision based model allows for a target dependent weighting of the relevant low level features but still does not predict the fixation locations as has been claimed based on more artificial tasks as well (Najemnik & Geisler, 2008). This failure argues against models in which top-down control of eye movements simply adjusts the weights of bottom-up features to guide eye movements (Itti & Koch, 2000; Treisman & Gelade, 1980; Wolfe, 1994), i.e. against top-down low-level control.

The reason why the low-level bottom-up influence on the chosen fixation locations here seems unimportant for the exploration of natural scenes might be that onsets or movements are necessary to make something salient in the sense of attracting fixations against top-down control (Jonides & Yantis, 1988; Yantis & Jonides, 1990). This would fit the pattern that these influences have some influence immediately after the sudden image onset, but not later. Also, the classical experiments, which show bottom-up driven saccades

use sudden onsets (Hallett, 1978) and some experiments with dynamic stimuli show that some dynamical objects attract almost all fixations (Dorr, Martinetz, Gegenfurtner, & Barth, 2010). Indeed, the main advantage of using low-level features for eye movement control seems to be a slightly faster response. For any stationary things this advantage may be irrelevant, which would give a normative reasoning for the restriction of low-level features governing eye movements to dynamic stimuli, especially as humans seem to be able to use high-level features quite early as well.

4.5.3 *Physiological substrate*

While I obviously cannot prove connections to physiology based only on behaviour, I can discuss whether the data I observe are compatible with our knowledge of physiology. The brain area most associated with the planning of saccades is the superior colliculus (White & Munoz, 2011). This structure was causally linked to the production of eye movements (Carello & Krauzlis, 2004) and contains a retinotopic map of visual inputs at the superior part and a corresponding map of saccade targets linking to motor areas in the inferior part. The superior part which receives input from many visual areas might be a realistic locus for the bottom-up influences on eye movements (White, Berg, et al., 2017; White, Kan, et al., 2017), while the inferior part, which receives input mainly from the frontal cortex through the frontal eye field may integrate the bottom-up influences with top-down control, although the frontal eye field does project to brain-stem areas bypassing the superior colliculus as well. If the superior part of the superior colliculus is indeed a representation of the bottom-up influences our findings that high-level information immediately contributes to eye movement control would be easily accommodated by stating that the influences of the higher visual areas are important for behaviour, not only the input from primary visual cortex, which should be captured by our early vision based saliency model. Similarly its easily conceivable that the top-down control may be strong enough to outvote the bottom-up influences when direct control is necessary.

4.5.4 *Future prospects*

I found that exploring the temporal dynamics of eye movement behaviour throughout a trial provides interesting insights into the control of eye movement behaviour even within a single fixation duration (Anderson et al., 2016; Rothkegel et al., 2017). These dynamics have been studied earlier (e.g. Over et al., 2007; Tatler & Vincent, 2008), but there are few models which produce dependencies between fixations at all (see Clarke, Stainer, et al. (2017); Engbert et al. (2015); Le Meur and Liu (2015) for notable exceptions) and even those who do are rarely evaluated regarding their abilities to produce natural dynamics and generally do not handle a connection to the explored images. Here I only scratch the surface of the possibilities to check models more thoroughly using the dynamics of eye movements. We now have the statistical methods (Barthelmé et al., 2013; Schütt et al., 2017) and datasets to pursue this research direction further and that the information from the dynamics of eye movements will be informative for better models of human eye movements including the connection to the explored images.

4.5.5 *Conclusion*

Investigating the temporal evolution of the fixation density over static images, I propose to separate three phases: An onset response, the main exploration and a final equilibrium state. Throughout the second and third phase fixation-locations are governed by similar high-level features and in visual search top-down control can almost completely overrule the bottom-up control. The only exception to these rules is the first saccade which targets a different density than later saccades, can be predicted by bottom-up models, to some degree even in visual search and contains interesting shifts over time within the single fixation duration, although high-level features improve predictions of fixation locations already $200ms$ after image onset.

5

DISCUSSION

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparametrization is often the mark of mediocracy.

Box (1976)

In this thesis I present an image computable early vision model to extract a realistic early vision representation of natural images. I use this early visual representation to construct a saliency map to predict fixation locations. Additionally, I present a thorough method to evaluate eye movement models successfully, even if they contain dynamic aspects and dependencies among fixations. Together these steps form a method to model scanpaths on arbitrary new images, which accomplishes the main goal of this thesis. However, many alternative approaches exist and the model I propose to predict eye movements in this thesis could be extended in many directions. In this chapter I want to discuss some of these possible additions and the implications of the model's successes and failures for some ongoing discussions.

5.1 EMBEDDING OF RESULTS

Overall my thesis lays substantial ground work for the development of dynamical eye movement models and succeeded in providing a dynamic eye movement model for previously unseen images.

5.1.1 *Early vision model*

My first step in Chapter 2 was the development of an image-computable early visual processing model to provide a realistic early visual representation to investigate the influence of early visual processing on eye movements. This model in itself is already a contribution as there had not been an image-computable model of early vision of this complexity and success before.

For the modelling of eye movements, the advantage of using an early vision model, which explains detection and discrimination data well, is twofold. First, explaining these data is evidence that the features I use truly represent early spatial visual processing adequately, going beyond the superficial resemblance other low-level features have to rely on. Second, fitting the discrimination data fixes the parameters of the model such that the eye movement data do not need to constrain the way the features are calculated originally. This reduction in free parameters simplifies the statistics substantially and reduces the amount of data necessary to fit the model. Or conversely, fewer parameters imply more accurate parameter estimates and less overfitting with the same amount of data.

5.1.2 *Evaluation methods for eye movement models*

My possibly largest contribution towards the development of better eye movement models so far are the evaluation methods for models I describe in Chapter 3. These methods are extremely general, as they provide a method for the complete statistical treatment of any causal eye movement model, which is able to predict a density for the next fixation location. This requirement should be fulfilled for any mechanistic eye movement model, as eye movement planning cannot be based on future events. This scope models to employ attention, high-level processing, sequential dependencies or even domain specific processing as for reading models.

These statistical methods for treating eye movement models were central for the analysing the temporal dynamics in Chapter 4. Furthermore, I used the methods to improve the SceneWalk model as I present in Chapter 3 and showed that dynamical aspects of eye

movements are indeed important for predicting them. Having these techniques available will facilitate the development of any future models aiming to predict whole scanpaths.

A similar movement towards likelihood based evaluation methods was started by Barthelmé et al. (2013) for models treating fixations as draws from a spatial point process. As saliency models are a special case of this model class these methods are directly applicable to those models (Kümmerer et al., 2015; Kümmerer, Wallis, & Bethge, 2017). For saliency models the main advantage of the likelihood approach is the unification of the different metrics previously used for evaluation.

5.1.3 *Connecting early visual processing to eye movements*

As I discuss in more detail in Chapter 4, low-level features—as represented by my early spatial vision model—do not provide a sufficient description of the guidance of eye movements. Higher-level information on objects in the scene as provided by a DNN improves predictions beyond the possibilities of low-level features. All bottom-up models with or without high-level features fail to explain eye movements in visual search (beyond the first saccade). Thus, one needs to include high-level guidance into a complete model of eye movement behaviour. As I also show that a simple reweighing of the low-level features is insufficient to explain the chosen fixation locations during visual search, including the high-level guidance might prove to be quite difficult.

At first glance these observations look like I manoeuvred myself into a dead end, as early visual processing seems to be a small factor in eye movement control. However, many ways lead out of this apparent dead end. Bottom-up control still seems to play a role for the first saccade (Chapter 4; Anderson et al., 2015). For this first fixation the evolution of the fixation density seems to be tightly coupled to the time since image onset (Chapter 4; Rothkegel et al., 2017). Thus, the sudden onset of the image seems to contribute to relatively low-level bottom-up eye movement control. Whenever things suddenly appear or move this bottom-up low-level control might be more important, warranting continued interest. Also, top-down control still has to rely on our perception of the world. Indeed, we are more sure that the restrictions on reportable perception are similar to those available for top-down eye movement control than we are for bottom-up control. Thus, modelling visual processing remains equally important for eye movement control although a focus on top-down and high-level factors means that we need to model the processing further than previously thought. Furthermore, I observed that the dynamics and systematic tendencies in eye movement control are important and informative, which leaves the study of these effects and their interaction with image dependent factors as another interesting way forward.

5.1.4 *Other studies I contributed to*

Beside the studies I present in detail in this thesis, I contributed to a number of additional studies lead by Lars Rothkegel in Potsdam. As some of these studies corroborate the conclusions I draw in this thesis, I want to shortly summarize the relevant findings here.

In a first study (Rothkegel et al., 2016), the starting position for the exploration of natural scenes was manipulated experimentally. When subjects had to start their exploration near the left or right border, their mean fixation location does not converge to the center of the screen gradually, but shows an overshoot towards the opposite side of the image, which lasts for several seconds. This tendency provides the main evidence for the

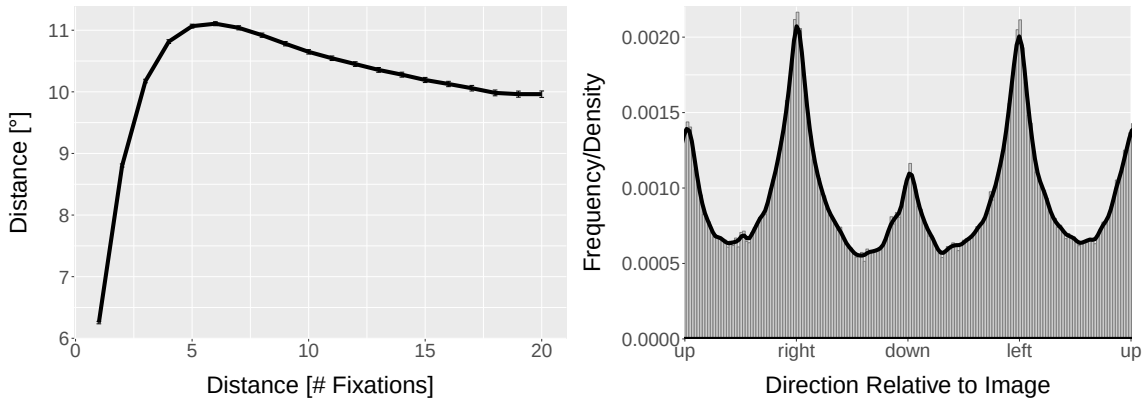

Figure 39: Two essential systematic tendencies in scene viewing, as measured in our scene viewing corpus. Left: Distance between fixations in degrees of visual angle plotted against the distance between fixations within the trial. Right: Histogram of Sakkade directions, which shows clear peakes around the cardinal directions, especially the horizonal axis.

inclusion of a relatively slow push away from previously visited locations as the inhibitory part of the SceneWalk model. Later we observed a similar overshoot in the distance to previously visited locations in the Corpus dataset as an average over all fixations (Fig. 39: left plot).

In a second study (Rothkegel et al., 2017), we investigated the initial central fixation bias in more detail. Here, participants were forced to keep fixating for a short time after image onset, resulting in a much weaker initial central fixation bias. This decrease in central fixation bias had a clear dependency on the saccade latency after image onset, which largely explained the effects of varying the enforced minimal fixation duration after image onset. These observations already allowed us to coarsely implement the initial center bias into the SceneWalk model and will certainly be important for a more detailed implementation in the future. Also the first fixation is the one for which low-level and/or bottom-up features might play a role. Thus, these results are of immediate interest for anyone interested in these low-level factors.

Finally, in our third study—the data of which I used in Chapter 4—we asked subjects to search for targets in natural scenes (Rothkegel et al., 2018). Besides the implications for saliency and the choices of fixation locations I describe in Chapter 4, we also find that the dynamics of the eye movements were adjusted to the target subjects searched for. Subjects made longer saccades and longer fixations for lower spatial frequency targets. This observation also supports the notion of relatively fine adjustment of the eye movements to the task. Furthermore, we observe, that the relatively short forward saccades display less guidance by the image content than other saccades. This suggests that there might be different contributions to the final choice of fixation location, which represents a strong starting point for improving dynamical eye movement models.

Furthermore, I helped supervising two bachelor candidates (Robert Geirhos and Carlos Medina-Theme) working on the stability of deep neural networks (DNNs) for object recognition against image distortions. This work, which we currently revise in response to reviewers' comments, shows that current DNN models of object recognition are still considerably less robust against distortions like noise on images. The main conclusion for the purposes of this thesis is that a model of high-level vision, which accurately represents what humans can and cannot see, has not been achieved by DNNs yet and might still take a while to be developed.

5.2 ALTERNATIVE APPROACHES

Our approach is not the only one towards understanding eye movement trajectories of human observers. To the contrary, eye movements are studied by many different researchers with different backgrounds, interests and aims and many different approaches have gained valuable insights into human eye movement behaviour and provide interesting information for models of this behaviour, which I discuss in this section.

5.2.1 *Other evaluation techniques*

Although moving towards a likelihood based evaluation of dynamical eye movement models is an important step forward for the evaluation of eye movement models, there are complementary statistical approaches for the evaluation of eye movement data.

The first alternative approach was already mentioned in Chapter 3: Besides model based evaluations, there is always an interest to find descriptive statistics to measure specific aspects of a behaviour. For eye movements these summary statistics include all measures of systematic tendencies (Tatler & Vincent, 2008, 2009), like saccade lengths (Tatler et al., 2006), spatial statistics (Barthelmé et al., 2013), angles between saccades (Smith & Henderson, 2009) and many more. These are complementary to the statistics we provide for the evaluation of models, which can only measure influences which are part of the model. As I describe in more detail in Chapter 3, these methods can be extended to provide a proper evaluation of models. For the final model evaluation, such pseudo-likelihood methods are usually inferior to the direct evaluation of a likelihood, when the correct likelihood is available, but can allow evaluations which focus on modelling specific aspects of the data, which might be desirable sometimes.

The second approach for evaluation of scanpaths is to transform the problem into a model for which inference methods are established, most commonly into a generalization of the linear model (Kliegl, Wei, Dambacher, Yan, & Zhou, 2011; Nuthmann, 2017; Nuthmann & Einhäuser, 2015; Nuthmann, Einhäuser, & Schütz, 2017). This approach has the advantage that both statistical (McCullagh & Nelder, 1989) and computational methods (Bates, Mächler, Bolker, & Walker, 2014) are well established for these models. Additionally, these models are commonly used for the analyses in related fields for the analysis of experimental data, such that results are well interpretable. This approach is especially interesting to test whether some factor influences eye movement control, without specifying a concrete mechanistic model. These approaches additionally allow the inclusion of image and/or subject as a random factor, moving further to generalized linear mixed models (Nuthmann et al., 2017), which is considerably more standardized than the detailed modelling of inter-subject differences in the likelihood approach I present, although Bayesian hierarchical models are often treated as the Bayesian analogue of mixed effect models (Gelman & Hill, 2006), when they are applied to generalized linear models. One drawback of fitting a generalized linear model is that this does not naturally contain a normalization step with the other locations. If factors increase the chance of some image area to be fixated, this does not automatically decrease the probability of other areas, although this is clearly true for fixations, as their number is limited. However, including such an influence would break diverse positive aspects of generalized linear models like the convexity of the underlying optimization problem. The other slight disadvantage of this approach is that it reduces the data to the question which parts of a grid were fixated. Nonetheless, this approach seems to be a versatile method to investigate whether some factor influences which areas of an image are fixated.

Eye movement models can also be analysed as spatial point process models (Barthelmé et al., 2013). For this kind of model, statistical (Baddeley, Rubak, & Turner, 2015; Illian, Penttinen, Stoyan, & Stoyan, 2008) and computational methods (Baddeley & Turner, 2005; Baddeley, Turner, Mateu, & Bevan, 2013; Illian, Sørbye, & Rue, 2012) are also established, although they are considerably more complex than for generalized linear models. The main advantage of spatial statistics is that they can take the spatial structure of the model into account, allowing dependencies between nearby spatial locations or smoothness assumptions for non-parametrically estimated influence factors. Also, these techniques are developed and justified without reference to a discretisation, such that the discretisations used for numerical calculations are only relevant for the quality of the numerical approximations, not for the statistical inference per se, similar as for my likelihood based model evaluations. Also, Barthelmé et al. (2013) point out spatial point process models can model a fixation density over time, which is an advantage as the fixation density does change over time (see Chapter 4). One drawback of spatial statistics models is, however, that they do not naturally include any notion of a scanpath, i.e. of an ordered set of fixations through time.

A third way of analysing scanpaths is based on regions of interest (Santella & DeCarlo, 2004). Analysing scanpaths based on regions of interest originally stems from research in which scenes are tailored to the research question, such that one is interested in the probabilities of subjects looking at specific regions of interest, like objects, features or separated regions of the scene. Examples for this research approach include face perception (Sammaknejad, Pouretemad, Eslahchi, Salahirad, & Alinejad, 2017) or playing games (Borji et al., 2014). If one can interpret the fixation and transition probabilities, these can serve as effective summary statistics about the scanpath. Also, some methods were developed to measure the difference between two scanpaths based on which regions were visited in which order (Le Meur & Baccino, 2013). As these metrics are usually only sensible when the scanpaths are quite similar, the success of these methods was variable. However, given the scarcity of similarity measures between whole scanpaths, there are no better methods to measure the similarity of scanpaths. To base the analysis of models on regions of interest, one can optimize models to predict the probabilities to fixate specific regions of interest, possibly including influences of the scanpath up to that fixation. In the limit, separating the image into a dense grid of pixel sized regions of interest is equivalent to my likelihood computation. However, statistical research towards the use of regions of interest for the evaluation of scanpath models rather fixated on finding good, informative separations of the scene into regions of interest (Privitera & Stark, 2000). Viewed from the likelihood perspective, this approach represents a form of data reduction and regions should be chosen to minimize the loss of information compared to the likelihood. This aspect might be interesting, as I am not aware of any work exploiting this connection to the likelihood for the choice of regions of interest.

As these alternative methods aim to test which factors influence eye movements and to summarize dependencies in the data and not to fit models and evaluate them, these methods may form one statistical toolbox with my my likelihood based model evaluation approach rather than competing with it. Using this improved toolbox may help the research community to improve our understanding of eye-movements and especially to build better models.

5.2.2 *Non-mechanistic models*

In this thesis I focussed on mechanistic models of eye movement behaviour, i.e. models whose components shall correspond to some mechanisms employed when humans choose where to look next. However, this correspondence is not necessary for simulation or prediction of eye movements and other researchers have proposed models which do not claim that the parts of the model correspond to some mechanism of eye movement selection. Instead the influence factors in these models usually correspond to observed statistical regularities in the data.

For example, Le Meur and Liu (2015) proposed such a model, which regenerates the saccade length and direction distribution by simply multiplying the predicted density with the measured saccade direction and length distribution. Later this approach was even extended by making the used saccade length and direction distribution dependent on the category of the scene (Le Meur & Coutrot, 2016).

Another model of this type was developed by Clarke, Stainer, et al. (2017) under the pretext of developing a better benchmark for the evaluation of eye movement models. Their "saccadic flow" model predicts the fixation locations independent of image content. To do so, it predicts a truncated Gaussian over the image, whose parameters were fit to smoothly depend on the location of the previous fixation.

The Clarke, Stainer, et al. (2017) model represents one important role of non-mechanistic models as a benchmark for the maximal possible prediction quality. Indeed it is unclear how well scanpaths can be predicted in general, as I referenced as the lack of a gold standard in Chapter 3. Thus, a measure how well eye movements can be predicted, given a relatively free mapping from a chosen set of predictors is of great interest for modellers who aim to design a more mechanistically plausible mapping for these predictors. Similarly, measures of the maximal prediction quality possible, while using some set of predictors can be informative for the question, which predictors are important and should be prioritised.

Also, for the further development of mechanistic models, models implementing an arbitrary statistical regularity in a simple way can be used to check whether aspects of eye movement behaviour are necessarily coupled or not. As an example of this use of statistical models, we used such a model in our paper on the influence of the initial fixation location (Rothkegel et al., 2016), to show that the relationship between subsequent saccade directions does not necessarily produce the overshoot to the opposite half of the image we observed.

Additional to these auxiliary functions for the development of mechanistic models, purely statistical models have a purpose in their own right in some contexts. In an applied context for example, the aim of a model might be mere prediction quality. In this context, mechanistic realism is irrelevant and if the statistical model makes better predictions it should be preferred. Also, statistical models can play an important role in an experimental or exploratory context, when researchers ask whether and how a particular parameter influences eye movement behaviour (Nuthmann, 2017). Such questions are important for understanding eye movements and usually have to be answered before a mechanistic model which reproduces the observed relationships can be developed.

Nonetheless, the ultimate goal for modelling eye movements is to produce a model which works similar to humans and not only reproduces some statistics. At some point mechanistic models might also produce decisively better predictions for eye movements and when this happens a purely statistical approach for the same relationship would be entirely obsolete. Until then, statistical models will continue to play an important

role, although their contribution to understanding eye movements rarely goes beyond the empirical findings.

5.2.3 *Non-image-computable approaches*

In this thesis I focussed on analyses and modelling based on features which can be computed from the image, potentially relying on other information as well. However, many aspects of image content cannot be computed reliably and validly from images so far and this hurdle has not diminished researchers interest in these factors like object positions (Einhäuser, Spain, & Perona, 2008), scene organization or viewer expectations (Henderson et al., 1999). To enable themselves to study influences on eye movements which cannot be extracted from the stimuli, researchers use various techniques.

The first possibility to test image properties one cannot automatically extract from images is to ask subjects to annotate images regarding the properties. Depending on the preferred analysis methods the annotations can be converted into regions of interest or features maps. This approach has been employed to study the influence of objects on fixation location (Einhäuser, Spain, & Perona, 2008; Stoll et al., 2015) as well as for the influence of expected target location (Mohr et al., 2016). The main drawback of this approach is that it remains an empirical question how reliable, valid and/or predictive the annotated properties are. If they are not predictive of fixation locations or not valid measures of the properties one is interested in, this might mislead conclusions about eye movement control, although this restriction applies to algorithmically extracted measures as well of course. Compared to image computable features human annotations additionally are costly and relatively coarse.

If one is not interested in why subjects look where in a scene, one can also avoid handling the image dependence by using the empirical density of fixation locations of different subjects. For example, we used this approach for the development and analysis of the SceneWalk model (Engbert et al., 2015; Schütt et al., 2017). Given enough data, this empirical density can function as a perfect saliency model. However, this approach does not contribute to the question where people look and ignores possible interactions between content and eye movement dynamics. Furthermore, it does not allow any easy generalization to new images. However, it is a viable option to concentrate on the dynamical aspects of eye movements.

Another similar approach is to infer regions of interest instead of a fixation density. The most common method here are cluster analyses, which are used to separate fixations into clusters, which are believed to correspond to objects or regions of interest in the scene (Santella & DeCarlo, 2004). Using these regions of interest, one can analyse the dynamics as with predefined regions of interest. A drawback is, that predictions for scenes require new accompanying eye movement data, defeating any purpose of the models for prediction.

These analyses based on other experimental data about the scenes are important tools to study aspects of eye movements which are not yet computable based on the stimuli. However, they usually result in qualitative statements like ”objects attract eye movements”, which require significant elaboration before they can be implemented into mechanistic, quantitative models.

5.2.4 *Search models*

There is a broad literature on visual search, which has clear connections to models of eye movements, because eye movements are often necessary to find a target if the search task is difficult (Hulleman & Olivers, 2017). Nonetheless, a substantial part of the visual search literature avoids eye movements by enforcing fixation, while subjects search for a target object among isolated distractor objects, regularly placed around the fixation spot (Müller & Krummenacher, 2006).

Continuing the tradition of using search displays with isolated target and distractor objects most research on visual search prefers such simplified displays even if eye movements are allowed or studied explicitly (Adeli et al., 2017; Zelinsky et al., 2013). The main connections analysed with this method are influences of target and distractor properties on the search time and the amount of errors, especially in connection to the number of distractor objects. The typical conclusion drawn from such studies is whether the search time depends on the number of distractors or not (serial vs. parallel search Wolfe & Horowitz, 2004).

The major models of visual search are build around Feature Integration Theory (Treisman & Gelade, 1980), which assumes that some feature maps are computed automatically and in parallel, while integrating information from multiple maps requires visual attention, which needs to be shifted serially over the objects in the display. The most prominent model following this tradition is Guided Search (Wolfe, 1994, 2007; Wolfe et al., 1989), which additionally allows the multiple simple maps to guide attention in search and can explain a large part of the visual search literature. Nonetheless, this perspective on visual search has recently been questioned as other researchers propose that the major factor governing search performance is the peripheral discriminability of the target from the distractors and thus, the size of the functional visual field, which is able to detect the target (Chang & Rosenholtz, 2016; Hughes, Southwell, Gilchrist, & Tolhurst, 2016; Hulleman & Olivers, 2017; Rosenholtz et al., 2012).

Essentially for this thesis, Guided Search and most other models developed for visual search focus on predicting the search time and error distribution results, but no eye movements (Hulleman & Olivers, 2017). This approach is sensible as most visual search studies do not record eye movements. For modelling eye movements however, these models are surprisingly uninformative.

One model, which does predict eye movements and even works on natural images is the model by Zelinsky (2008). In overall form this model is similar to the one I present in this thesis. It generates a visual representation of the image which is compared to the representation of the target to generate a priority map for the selection of the next fixation location. The model however contains a representation of superior colliculus' processing, which smooths the priority map depending on the current fixation location. This process can explain some previously puzzling observations like fixations which fall between objects (Zelinsky, 2012). More recently, this model has been extended to allow a categorical specification of the target and to model free viewing data as well (Adeli et al., 2017; Zelinsky et al., 2013). These modelling results suggest adding a model of the superior colliculus processing to our model, which could be a step for future modelling.

5.2.5 *Fixation durations*

In this thesis I focus on fixation locations. Another approach focusses on the modelling of fixation durations. Fixation durations were studied in high detail for specialized tasks such

as reading (Engbert et al., 2005; Reichle et al., 2003) and in highly controlled laboratory tasks as saccadic reaction times (Colonius & Diederich, 2004). For fixation durations in these specialized tasks race and diffusion models (Ratcliff, 1978; Ratcliff & Smith, 2004) provide good fits to the data, as for reaction times in general. Therefore, more complex variations and combinations of such models were proposed as models for reading (Engbert et al., 2005), eye movements in general tasks (Trukenbrod & Engbert, 2014) and also for eye movements in natural scenes (Nuthmann et al., 2010). The models for fixation durations in natural scenes (Nuthmann, 2017) study the same influence factors as the ones for fixation locations (bottom-up, top-down & systematic tendencies). These models typically allow more complex interactions among influence factors and the dynamic naturally plays a more important role.

Recently, models of fixation duration were connected to models of fixation location (Einhäuser & Nuthmann, 2016; Tatler et al., 2017). The central observation is that models which are designed to predict where people look fast also predict were subjects look at all. This especially includes dynamical factors such as the saccade direction or angles between saccades, but also image factors such as the contrast or salience at the target location. These (correlative) connections between the control of fixation duration and fixation location suggest that a combined model which fits both fixation durations and locations simultaneously might be possible. The most common ideas to implement such a model either propose a race of possible locations for the selection as the target (Tatler et al., 2017) or a central timer based on a single diffusion process (Trukenbrod & Engbert, 2014). Such a model would be advantageous as it could use more of the information present in the data. Especially, a combined model for fixation durations and fixation locations could allow researchers to pinpoint effects which can be isolated only in one of the measures.

5.2.6 *Attention models*

I do not include explicit attention effects in any of the models, although attention could influence both in the visual processing I model in Chapter 2 (Schütt & Wichmann, 2017) and on the selection of eye movements we model with the SceneWalk model (Engbert et al., 2015).

The interaction of eye movements and attention is not clearly unidirectional. While attention and eye movements are certainly related (Deubel & Schneider, 1996), it is unclear whether eye movements enforce the deployment of attention or whether eye movements follow attention. Nonetheless, these interactions would be interesting for the modelling of eye movements if they explained some experimental observations on eye movements. Unfortunately, the only connection of this kind I know of is inhibition of return, whose influence in natural scene viewing is still debated (Hooge, Over, van Wezel, & Frens, 2005; Rothkegel et al., 2016; Smith & Henderson, 2009). Whether more experimental observations on eye movements are mediated by attention effects is a topic for further research.

Another connection between eye movements and attention is that researchers on attention sometimes subsume eye movements under attention as "overt attention shifts". Thus, attention models should incorporate eye movements (Hulleman & Olivers, 2017). Indeed attention models are structurally similar to usual models of higher level perception, but include top-down influences and/or lateral interactions. Prominent variants are biased competition (Desimone, 1998) and selective tuning (Tsotsos et al., 1995). Thus, ideas how to model the influence of attention on processing are available. For my models how-

ever, these ideas were not relevant, because my early spatial vision model does not contain any higher level processing and thus cannot include any top-down influences.

Nonetheless, models of attention could provide eye movement models. Most attention models are neither image-computable nor make explicit predictions about eye movements. However, at least two notable exceptions to this trend exist: The models by Borji et al. (2014) and Wloka, Kotseruba, and Tsotsos (2017). The model by Borji et al. (2014) is originally based on the groups' earlier saliency models (Itti et al., 1998) and added the influence of task as attention effects. In contrast, the model by Tsotsos et al. (1995) was long developed as a pure attention model and was only recently extended to predict eye movements (Wloka et al., 2017). However, both models postulate a hierarchy of visual processing, whose processing can be influenced by top down attention and whose activities are finally combined into a priority map for the selection of the next target of visual attention. Unfortunately the model by Borji et al. (2014) was not evaluated regarding eye movement dynamics, such that I cannot comment on its performance as an eye movement model. The model by Wloka et al. (2017) was indeed evaluated regarding eye movement dynamics but makes wrong predictions for the distribution of saccade amplitudes already such that this model cannot be taken seriously as a model of eye movement dynamics yet. Nonetheless, it would be interesting to see how these models fare in explaining eye movements, as could be analysed with the methods I present in Chapter 3 (Schütt et al., 2017).

5.3 CONTROVERSIES

Although I focussed on the implementation and evaluation of models in this thesis, some of the choices I made reflect opinions about long standing controversies and some of the experimental observations we made have some bearing on these controversies I want to discuss in this section.

5.3.1 *Automated vs. cognitive control*

The first controversy I want to discuss is how automated the control of eye movements is or conversely how close the cognitive control over eye movements is. This controversy has been discussed for a long time (Hallett, 1978) and is related to the question whether eye movements are controlled by top-down or bottom-up processes, which I discussed in Chapter 4, concluding that eye movements contain some bottom-up control initially, which quickly incorporates high-level features and can be overruled by top-down control. The question how automated eye movement control is, has a somewhat different slant, asking whether eye movement control is cognitively penetrable and whether it requires cognitive resources rather than which influence factors are important.

In favour of an automated eye movement control, there are the following findings: First, there are cases of clear exogenous overt attention, i.e. some stimuli attract attention and eye movements even when subjects are instructed to ignore these distractions (Hallett, 1978). Second, when subjects explore scenes many systematic tendencies are observed (Tatler & Vincent, 2008, 2009). Some of these might be correlates of higher level control strategies, but generally they are explained well as a signature of the underlying control mechanism. Third, subjects' awareness of their own eye movements seems to be limited (Clarke, Mahon, Irvine, & Hunt, 2017; Kok, Aizenman, Võ, & Wolfe, 2017; Võ, Aizenman, & Wolfe, 2016), which makes a cognitive control of individual eye movements unlikely for

usual circumstances. Indeed, many people are not even aware that their eye movements consist of fixations separated by saccades, and saccades were first described towards the end of the 19th century (summarized in the history of eye movement research by Wade & Tatler, 2011), which would be surprisingly late if we were generally aware of our eye movements.

In favour of cognitive control of eye movements, there are the following findings: First, tasks and context can influence eye movement control as has been long known for fixation locations (Castelhano et al., 2009; Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007; Land & Lee, 1994; Land et al., 1999; Yarbus, 1967) and is also true for other eye movement parameters as we showed in our visual search paper (Rothkegel et al., 2018). Second, as I mentioned in Chapter 4, predicting the choice of fixation locations requires the inclusion of high-level features of the scene like object locations (Einhäuser, Spain, & Perona, 2008; Nyström & Holmqvist, 2008; Stoll et al., 2015), which implies that eye movement control has access to high-level representations. Third, there is a wealth of research in which researchers ask subjects to fixate a specific location. While subjects are usually not perfectly obeying this instruction, the instruction certainly has an effect. One example of such an instruction in our own research are our investigations on the central fixation bias (Rothkegel et al., 2017), for which we asked subjects to keep fixating the initial fixation cross until it disappeared a while after the scene appeared.

Based on these evidence and the observations I made in Chapter 3, I conclude that eye movement control incorporates high-level information and can be guided by cognition, but is still highly automated, such that one may decide where one wants to look, but passes this decision to an automated process which generates the necessary eye movements independently. In my opinion this view nicely explains how cognition can influence eye movements without exerting direct control, that it takes effort not to make eye movements, that cognitive control fails sometimes and that subjects have little awareness of the eye movements they make.

Despite my acknowledgement that cognition may have an important influence on eye movement control, the models I present in this thesis do not explicitly contain any cognitive influences. The two reasons why I chose not to implement such influences are: First, general cognition is not well understood and thus including it satisfactorily into a model is practically impossible. While we can separate some broad parts of cognition, I am not aware of any model which could describe cognitive processing of natural scenes. Second, I genuinely believe that the two ends I model here are largely independent of cognitive control and can be studied independent of cognition. It is still debatable, whether cognition may penetrate perception at all (Firestone & Scholl, 2016) and if cognition affects perception the effects appear to be small. Similarly, the final automated part of eye movement control I argued for in the last paragraph can be studied independently, especially by analysing the dynamics and dependencies we observe in scanpaths.

5.3.2 *Inhibition of return*

One central mechanism in the SceneWalk model, which I employ in this thesis, is an inhibitory tagging of previously visited fixation locations (see Chapter 3 and: Engbert et al., 2015; Schütt et al., 2017). Whether such an inhibition of return plays a role in eye movement control and especially whether it influences the choice of fixation location or only slows eye movements returning to the previously visited location is controversial.

The concept of inhibition of return originally comes from the attention literature (R. M. Klein, 2000; Posner & Cohen, 1984). There, researchers observed an interval

a couple of hundred milliseconds after an orienting cue, when subjects are slower to react to a target appearing at the cued location than without any cue. This was interpreted as an inhibition of the return of attention to a recently visited location.

Based on the coupling between attention and eye movements, inhibition of return was soon added to models of eye movement control as well (Itti & Koch, 2000; Tsotsos et al., 1995) and is a central part for models of visual search to facilitate the exploration of the whole display (R. M. Klein & MacInnes, 1999; Zelinsky, 2008).

The idea was criticised soon (Hooge et al., 2005; Smith & Henderson, 2009), because subjects return to the previously fixated location more frequently than expected from random choice instead of less often as inhibition of return predicts, although these saccades are indeed delayed compared to other saccades. Both systematic tendencies are present in our data as well (Rothkegel et al., 2016).

However, the distance of the current fixation location from an earlier fixation location peaks at a distance of 5-6 fixations (a bit less than 2 seconds, see Figure 39). This observation, which we originally made for the first fixation only (Rothkegel et al., 2016) seems to imply a repellent influence of earlier fixation locations. As this effect is much slower than the tendency to return to the previously fixated location immediately, I believe these observations represent two separate compatible effects. The SceneWalk model does only implement the slow inhibitory effect and thus can only reproduce the overshoot in distance between fixations not the tendency to return to the previously fixated location immediately (Rothkegel et al., 2016).

5.3.3 *Maps vs. Objects*

In the models I describe in this thesis all processing is based on continuous activity maps. Such activation maps certainly are an adequate description for both the first retinotopic steps of visual processing and the finally chosen fixations and their densities. However, some research suggests that processing at some intermediate stage is object centred. For example, there is evidence for attention spreading along objects rather than between them (Roelfsema, Lamme, & Spekreijse, 1998; Theeuwes, Mathôt, & Kingstone, 2010; Vecera & Farah, 1994), some models of eye movements make successful predictions based only on the position of objects in scenes (Einhäuser, Rutishauser, & Koch, 2008; Stoll et al., 2015) and models of visual search and attention classically operate on separated objects (Treisman & Gelade, 1980; Wolfe et al., 1989), although this view has been challenged recently (Hulleman & Olivers, 2017).

I do not want to deny the possibility that some internal processing is centred on objects. Although the mapping I employ from early vision input to fixation density operates on activation maps all the way through, this mapping at best represents a small part of the guidance of eye movements and the paths which imply more thorough processing of the visual input might well incorporate object based processing.

Similarly, I do not have strong opinions on the reference frame of the maps I employ. In my models everything is referenced relative to the stimulus as this was easiest to implement in a model, which needs to predict fixations relative to the stimulus position. Realistically, both early visual processing (Engel, Glover, & Wandell, 1997) and the last steps of eye movement control (White & Munoz, 2011) are retinotopic. Indeed this retinotopy might be important and it is an interesting future aim to include the processing differences between different retinal locations into models as the ones I present here. For eye movement control one could make the blurring of the predictions location dependent, as has been done successfully for a different model (Zelinsky, 2012). Peripheral visual

processing processing is a large field on its own (Strasburger et al., 2011), providing a wealth of effects one could try to include as I discuss below. For higher levels of processing however, other egocentric or allocentric reference frames might be used. Possibly caused by our limited understanding of high-level processing in general it remains unclear whether this processing with other reference frames has any influence on eye movement control.

5.4 DIRECTIONS FOR FUTURE RESEARCH

In this section I discuss diverse possible model extensions and possibilities for future research to improve our understanding of eye movement control and its interaction with perception.

5.4.1 *Peripheral processing*

As a first model extension, one could include a peripheral decline of visual processing. The early vision model that I present implements no peripheral decline and thus handles all input with the full processing detail, which realistically is only available foveally. As the main purpose of eye movements is to move the fovea to interesting locations, it must rely on peripheral vision to select fixation locations. Thus, constraints of peripheral vision might be the most interesting perceptual constraints on eye movement control and it would be an important step forward to have a realistic model of peripheral visual processing.

For early visual processing the inclusion of a peripheral decline might be possible relatively soon, as the decline in simple capabilities like contrast detection and discrimination was measured extensively (Baldwin et al., 2012; Foley, Varadharajan, Koh, & Farias, 2007; Pointer & Hess, 1989; Rovamo, Franssila, & Näsänen, 1992; Rovamo & Virsu, 1979). Similarly, the corresponding physiological measurements exist of sampling densities in the retina (Curcio & Allen, 1990; Curcio, Sloan, Kalina, & Hendrickson, 1990; Dacey & Petersen, 1992) and the cortical magnification factors towards the periphery (Duncan & Boynton, 2003; Harvey & Dumoulin, 2011; Rovamo & Virsu, 1979; Virsu & Rovamo, 1979). Thus, the data base for an early vision model of the periphery is available. Furthermore, some existing models already include a peripheral decline (Bradley et al., 2014). Thus, an extension of my early spatial vision model to include a peripheral decline might be feasible, although it is not trivial as some aspects seem to differ between foveal and peripheral processing beyond receptive field sizes. For example, the normalization seems to have different parameters in the periphery (Xing & Heeger, 2000).

For eye movement models a peripheral decline would be especially interesting if it provided an explanation for some of the dynamical aspects of eye movements or restricted the possibilities what information is available peripherally to choose the next fixation locations. If one had a reasonably fast model of visual processing including a peripheral decline one could replace the currently used early vision model immediately and test whether the inclusion of a peripheral decline helps predictions and/or whether parts we currently model as an attention window or general bias for the choice of eye movements can be dropped.

5.4.2 *Higher-level processing*

When modelling the connection between early vision and eye movements in Chapter 4, I note that high-level properties of images explain eye movements better than the direct mapping from early vision features I tested confirming earlier reports that high-level features are import for eye movement control (Anderson et al., 2016, 2015; Einhäuser, Rutishauser, & Koch, 2008; Einhäuser, Spain, & Perona, 2008; Henderson et al., 2007; Judd et al., 2009; Kümmerer et al., 2016; Stoll et al., 2015; Torralba et al., 2006).

Thus, including higher level processing is another avenue for future research. Indeed, extending our understanding of higher level visual processing is currently an active field of research. There are some studies on human object recognition (Wichmann et al., 2006, 2010) and scene processing (Schyns & Oliva, 1994; Wichmann et al., 2010), which provide some insights what information might be used to perform higher level tasks and some conceptual models how mid- and high-level vision might work exist for a while now (Riesenhuber & Poggio, 1999). However, these models could never represent high-level vision restrictions of humans adequately.

For models of higher visual processing, the currently most promising framework are deep neural network (DNN) models (Kriegeskorte, 2015; D. L. Yamins & DiCarlo, 2016). These models are highly efficient implementations of multiple layers of simplified model neurons, containing only a linear weighting and nonlinearity. When these models are trained—i.e. have their weights optimized—to map images to object category (Russakovsky et al., 2015) or other high-level features (Zhou et al., 2017), they solve these tasks better than any previous machine learning algorithm. Additionally, the representations extracted by deeper layers of the networks are useful for other tasks as well, which has led to the solution for many other tasks for which data are not sufficient to train deep neural networks from scratch (Huh, Agrawal, & Efros, 2016; Yosinski, Clune, Bengio, & Lipson, 2014). This method also enabled the creation of the last generation of saliency models, which indeed predict the fixation density much better than earlier ones (see Chapter 4 & Huang et al., 2015; Kruthiventi et al., 2015; Kümmerer et al., 2016; Pan et al., 2017). Additionally, there is some evidence that these trained networks create similar representations as the human ventral stream (Cadieu, 2014; Kriegeskorte, 2015; D. L. Yamins & DiCarlo, 2016; D. L. K. Yamins et al., 2014).

However, DNNs react dissimilar from humans to image distortions (Dodge & Karam, 2016, 2017; Geirhos et al., 2017), which are usually used to implement the limitations of early visual processing or the peripheral decline like noise, filtering or eidolon distortions (Koenderink, Valsecchi, Doorn, Wagemans, & Gegenfurtner, 2017). Therefore, DNNs are currently not suitable to model the limitations of the visual system and distorting the input by some mechanism mimicking the constraints of early visual processing is unlikely to model constraints for higher level processing adequately. However, there are no other good image-computable models of the limitations of higher level visual processing either, and our understanding of these processes is limited in general. At best, there are some summary statistics based models (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz, 2016), which are reasonably successful in modelling visual processing one step further.

5.4.3 *Crowding: Peripheral restriction on higher levels*

Combining the arguments for high-level models and models of peripheral processing, the ideal vision model as a basis for eye movement models should provide a good model of high-level processing in the periphery.

However, higher-level peripheral vision seems to differ in important ways from foveal vision (Strasburger et al., 2011). The peripheral decline happens at many levels in the visual hierarchy, obscuring which decline is the limiting for eye movement control or any other task. Beyond the decline in sampling density in the retina, we know from neuronal and psychophysical data that there are further limitations at higher processing levels. In physiological measurements researchers observe that smaller proportions of the neuronal hardware are devoted to peripheral processing the higher one moves in the visual hierarchy and that peripheral receptive fields grow disproportionately faster than the central ones (Gattass, Gross, & Sandell, 1981; Gattass et al., 1988). In psychophysical experiments this seems to correspond to the observation of crowding (J. Freeman & Simoncelli, 2011; J. Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Wallis & Bex, 2012; Whitney & Levi, 2011), where nearby similar objects are detrimental for the recognition of objects in the periphery.

To explain mid-level peripheral processing, the currently most successful models are build around the idea of summery statistics (Balas et al., 2009; J. Freeman & Simoncelli, 2011; J. Freeman et al., 2013; Keshvari & Rosenholtz, 2016; Rosenholtz, 2016; Rosenholtz et al., 2012), i.e. the idea that higher levels of processing in the periphery represent a summary over space of some image features instead of a fully resolved map. Because of the similarity of this idea to the averaging of features done for the generation of textures (Portilla & Simoncelli, 2000), this kind of representation has been described as texture-like (J. Freeman & Simoncelli, 2011) and indeed, early attempts at using this idea for handling peripheral vision used features from a texture generation algorithm (Balas et al., 2009; J. Freeman & Simoncelli, 2011). By now, the features to be summarized have been extracted from deep neural networks as well, for both texture generation (Gatys, Ecker, & Bethge, 2015; Ustyuzhaninov, Brendel, Gatys, & Bethge, 2017) and models of peripheral processing (Wallis et al., 2017). However, this description of the visual periphery apparently covers the processing up to V2 rather than up to IT as DNNs claim for foveal processing. One advantage these summary statistics models have is, that they are image-computable, i.e. experiments and theoretical considerations can be based on images already. Relating the summery statistics model of crowding to eye movement control, one early study directly linked summary statistics to visual search performance (Rosenholtz et al., 2012). Also for visual search, it was shown that peripheral discriminability predicts performance better than foveal discriminability (Hughes et al., 2016)

In physiological terms these models aim to explain V2 (J. Freeman et al., 2013), in psychophysical terms they try to explain crowding effects (J. Freeman & Simoncelli, 2011). Beyond V2, we still have physiological observations (Strasburger et al., 2011), which provide a rough understanding what information is processed where in the brain, but our understanding how the processing works and consequently what its general and peripheral limitations are is limited.

Summarising the last three sections, one way of improving the models I present in this thesis is to extend the model of visual processing towards higher level processing and/or the periphery. Both these research questions are currently pursued. When this research yields new results they will allow further progress on the subjects I investigate here. However, researchers work on both these questions for decades by now such that a complete image-computable representation of human peripheral high-level vision is unlikely to emerge soon. Thus, eye movement models should use the progress as it is made instead of waiting for the elusive ultimate visual perception model.

5.4.4 *Eye movement dynamics*

SceneWalk—the model of the eye movement dynamics I employ in this thesis—is a strong simplification. It has only 2 mechanisms implementing dependencies between fixations. One mechanism—the attention window—attenuates the fixation density far away from the fixation location, which results in a reasonable distribution of saccade lengths in model simulations. The other mechanism implements an inhibition of return, which favours exploration and regenerates the distances of fixations which are separated by more than one saccade.

Still, the SceneWalk model can already produce much more realistic scanpaths than simple draws from the saliency map. However, there are many more known dependencies between saccades and fixations (Tatler & Vincent, 2008), as I can confirm based on the free viewing corpus I presented in Chapter 4. For example, subjects clearly prefer saccades along the cardinal directions of the explored image (Foulsham et al., 2008, see also Figure 39). Subjects also have a tendency to either continue in the direction they were moving in the previous saccade (*saccadic momentum*; Smith & Henderson, 2009; Wilming et al., 2013) or to return to the location they came from ("facilitation of return"; Hooge et al., 2005; Smith & Henderson, 2009). Furthermore there are some tendencies which develop over the course of a trial: For example, there is a strong central fixation bias at the start of the exploration (Tatler, 2007), whose exact temporal evolution we explored in a recent study (Rothkegel et al., 2017). Another example is the *coarse-to-fine-strategy*, i.e. the observation that saccade lengths decrease over the course of a trial while fixation durations increase, at least in search tasks (Over et al., 2007; Rothkegel et al., 2018).

To further pursue the goals of the SceneWalk as a mechanistically realistic model of eye movement dynamics, these dependencies should not be implemented as a simple statistical building block, but as a realistic mechanism. Such mechanisms to explain the observed dependencies seem to be relatively easy to conceive though. Indeed we present an implementation of the initial central fixation bias in our paper on the temporal evolution of this bias (Rothkegel et al., 2017). For the dependencies between successive saccades we have at least collected substantial information in our article on visual search (Rothkegel et al., 2018), where we argue that our observations are compatible with a proportion of relatively unguided saccades which explain the saccadic momentum. Implementing such additions to the SceneWalk model would be interesting as a further direction for the modelling of eye movements.

Including more dependencies into models is not only of immediate interest to improve the models, but the performance of different mechanisms in explaining eye movements may be informative for diverse discussions about the purpose and implementation of eye movement control. For example, there is an ongoing discussion about the relevance of return saccades and interaction with inhibition of return (Hooge et al., 2005; Smith & Henderson, 2009; Wilming et al., 2013). Here, an implementation including both mechanism at their correct timescales in a model might be of great interest to investigate whether these mechanisms can coexist, how they interact with image borders (Wilming et al., 2013) and how well they explain the observed dependencies between saccade directions.

Another reason to pursue this direction is that few mechanistic models of eye movement dynamics in natural scenes exist so far, such that many questions remain to be answered. This observation contrasts to the state for eye movement models in reading (Engbert et al., 2005; Reichle et al., 1998) or models of early visual processing (Bradley et al., 2014; Foley, 1994; Schütt & Wichmann, 2017; Teo & Heeger, 1994). One reason why so few models were explored so far may be that the evaluation of such models was considered

hard. As I presented a comprehensive set of tools to fit and evaluate models of dynamical eye movement models (Chapter 3; Schütt et al., 2017), such concerns should be largely overcome.

Slightly further into the future a model of eye movement dynamics should most likely model fixation durations as well. Fixation durations were traditionally analysed separately from fixation locations (Nuthmann et al., 2010; Trukenbrod & Engbert, 2014). Recently however, there is an increased interest in combining the two (Einhäuser & Nuthmann, 2016; Tatler et al., 2017), as researchers observe that many influences are shared between fixation duration and location selection. Thus, a combined modelling of fixation locations and fixation durations might benefit models of both measures, especially when the strong dynamics and dependencies between fixations are taken into account, because these dependencies like angles between saccades seem to influence both measures simultaneously (Over et al., 2007; Tatler & Vincent, 2008).

5.4.5 *Dynamics and image content*

One interesting aspect of the dynamics of eye movement behaviour is that they seem to interact with the image content and the fixation probability over the image. For example, the angle between successive saccades correlates with the fixation density and the saliency model prediction for the next fixation location (Rothkegel et al., 2018). At first glance these relationships might sound like a further complication. However, they might provide a handle to better understand how the different influence factors governing eye movements are combined to make a final decision where to fixate when. If some dynamical aspects of eye movements like fixation durations or saccade lengths and directions favour different influence factors, these separations might allow conclusions about the timecourse of the different factors and provide some data to constrain the decision mechanisms for eye movement control. These decision mechanisms are otherwise hard to constrain at least beyond whether the influences interact at all or not.

As a concrete research project one might search for dynamical aspects which identify groups of saccades which depend on some of the factors particularly strongly or particularly weakly. Such statistical regularities could then be tested for causality in experiments and might provide central evidence about the mechanism making the decision where to look next.

5.4.6 *Statistical improvements*

Last but not least, one way forward for the topics I discuss in this thesis is to improve the statistical methods. Indeed my Chapter 3 represents a purely statistical, methodological advance for the treatment of eye movement models. However, I still used relatively simple mathematical tools to solve the optimization and sampling problems I encountered for the eye movement and early vision models, which was sufficient for the models I used in this thesis. Ultimately the complexity of models is limited by the available statistical methods though and some of the calculations I present in this thesis took weeks and months to run. Therefore, more efficient methods and making them available more easily to researchers on cognitive models, might save large amounts of time for these researchers and might even allow the development of some complex models which cannot be fitted or evaluated today.

Concretely, there are three areas where better algorithms might aid the development of models like the ones I present here: *Optimization*, *sampling* and *analyses over time*. Indeed there is statistical research on all three of these, such that there are methods waiting for adjustments and applications for models of human behaviour.

Optimization is used for statistical purposes to find the best parameters for models or to find stimuli with optimal properties. In both cases the optimization problems can become complex quickly, because image stimuli are high-dimensional and because the models implement almost arbitrarily complex functions to determine their predictions. In principle, optimization algorithms are available in higher level processing languages like MATLAB or Python. For functions which are computationally expensive to evaluate it can be interesting to search for more specialised optimization techniques as there is active research about meta-modelling approaches, which try to fit the function to be optimized locally allowing a more sophisticated choice of points to evaluate (Acerbi & Ma, 2017; Hennig & Schuler, 2012; Hernández-Lobato, Hoffman, & Ghahramani, 2014). Similarly, for big datasets methods which allow the splitting of the dataset might be interesting (Bottou, 2010; Kingma & Ba, 2014; Zinkevich, Weimer, Li, & Smola, 2010).

Sampling methods are mainly used to approximate the posterior distributions when Bayesian analyses are used. As concrete improvements beyond the methods I showed, one could try to use Hamiltonian Monte Carlo methods, which allow the use of derivatives for better sampling efficiency (Duane et al., 1987; Hoffman & Gelman, 2014; Neal, 2011), sampling methods used for big datasets specifically (Korattikara, Chen, & Welling, 2014; Welling & Teh, 2011) or adaptive sampling algorithms which reduce the necessity to adjust the algorithms to the problem by hand (Haario et al., 2006; Hoffman & Gelman, 2014; Wang, Mohamed, & De Freitas, 2013) .

Anaylses over time refer to the need for specialized methods to handle the sequential nature of eye movement models. Here the overarching mathematical concepts are sequential Monte Carlo methods for sampling (Doucet et al., 2001) and more generally data assimilation which refers to methods which allow adjustments of sequential models to data corresponding to different points in time (Reich & Cotter, 2015).

Finally, models of eye movements might also profit from a more detailed mathematical analysis of the data, especially to generate other interesting summary statistics to provide ideas what might be missing in current models of eye movements. One promising candidate from mathematics to find such new summaries might be spatial statics, which was already used with some success on eye movement data (Barthelmé et al., 2013).

5.5 CONCLUSION

In this thesis I developed an image-computable early vision model, improved the evaluation methods for dynamical eye movement models and provide a model and careful analyses for the low-level bottom-up influence on eye movement control over time. Certainly, many aspects of the models could and will be improved by other researchers as I discussed in this last chapter. Nonetheless, this thesis provides important methods and a first extendible model, which already combines our knowledge about perception and eye movements.

A MATHEMATICAL DETAILS

The treatment of this paper has been rather qualitative and the evidence [...] is convincing more to intuition than to intellect. Nevertheless, the results will yield a little to analysis.

Naka and Rushton (1966)

For the analyses in Chapter 2 I require some technicalities, which were placed in an appendix for the original publication (Schütt & Wichmann, 2017). These appendices are repeated here largely as they were published. The first appendix describes the methods I used for fitting the model, in the second I calculate the derivatives of the model to the parameters and in the third I describe my techniques to optimize stimuli to create maximally differentiating stimuli.

A.1 FITTING

In the main text my presentation follows the order in which the experimental findings depend on each other, building up from grating detection experiments to masking by arbitrary natural images. As this is not the order in which I fitted the parameters to data, I explain the setting of parameters in this appendix in the order I fixed the parameters.

As our model computes a percent correct pc_i for any pair of stimuli to be discriminated, we can compute the likelihood L—the probability of observing the data given the model parameters—directly from the observed number of correct trials k_i and the total number of trials n_i in each specific experimental condition using the Binomial distribution $\mathcal{B}$:

$$L(\theta|\,\text{data}) = P_M(\text{data}\,|\theta) = \prod_{i=1}^{N} \mathcal{B}(k_i|n_i, pc_i) \tag{33}$$

$$= \prod_{i=1}^{N} \binom{n_i}{k_i} (pc_i)^{k_i} (1-pc_i)^{n_i-k_i} \tag{34}$$

As it is usually done we computed the log-likelihood l from this and removed constant factors from the equation:

$$l(\theta|\,\text{data}) = \log(L(\theta|\,\text{data})) = \sum_{i=1}^{N} \log\left(\mathcal{B}(k_i|n_i, pc_i)\right) \tag{35}$$

$$= \sum_{i=1}^{N} \log\left(\binom{n_i}{k_i}\right) + \sum_{i=1}^{N} \left(k_i \log(pc_i) + (n_i - k_i)\log(1-pc_i)\right) \tag{36}$$

$$= C + \sum_{i=1}^{N} \left(k_i \log(pc_i) + (n_i - k_i)\log(1-pc_i)\right) \tag{37}$$

For this log-likelihood we calculated a gradient over the parameters of the nonlinearity as detailed in Appendix A.2 and optimized using a BFGS algorithm as implemented in MATLAB's "fminunc" function.

As not all data are constrained in each condition, for which we needed a separate parameter fit, we had to fit the parameters in a successive fashion.

We first fixed the parameters of the preprocessing for all presentation times. For the optical distortions we fixed the pupil size to $4mm$ diameter as a rough estimate for the environment of psychophysical measurements. Using preliminary parameter estimates from the literature, we then fixed the initial neural weighting of spatial frequencies to fit the detection data for each presentation duration.

Next we fixed the parameters of the log-gabor-decomposition. We set the filter bandwidths to 40° and 1.4 octaves for orientation and frequency respectively based on rough estimates from earlier measurements. We then set the range of spatial frequencies to

Table 3: Parameter values evaluated in the grid search for parameters. For each parameter combination an optimal factor to the final variance was fit as a final noise factor.

Parameter	Levels
p	$1.00, 1.50, 1.60, 1.70, 1.80, 1.90$
	$2.00, 2.10, 2.20, 2.30, 2.40, 2.50, 2.60, 2.70, 2.80, 2.90$
	$3.00, 3.25, 3.50, 3.75, 4.00, 4.50, 5.00$
q	$0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$
$C \times 10^2$	$0.1000, 0.1292, 0.1668, 0.2154, 0.2783$
	$0.3594, 0.4642, 0.5995, 0.7743, 1.0000$
ω_θ	$0, 0.1112, 0.2594, 0.4447, \infty$
$\frac{N_F}{N_C}$	$0, 0.1, 1, 10$

$.5 - 20\frac{cyc}{deg}$ roughly covering the visible range of frequencies. For the number of channels we set the model to use 8 orientations and 12 frequencies to reduce the ripple artefacts in the output to a bearable range as described in the main text. At this stage, we also fixed the bandwidth of the normalization pool, setting the standard deviation of the Gaussian to be $\sigma_F = .5$ octaves.

Next we fixed the bandwidth of the normalization pool in orientation based on the oblique masking data for the $1497ms$ presentation time, for which we had most data. To do so we computed the likelihood for a grid of parameter values over the normalization bandwidth ω_θ, p, q and C.

One computational trick we used to reduce the number of parameters to evaluate was to fit the overall noise variance independently of the other parameters. This can be done very efficiently, because scaling the noise for all pixels and all channels by the same factor c_ϵ does not change the optimal decoding scheme. Thus the signal to noise ratio (SNR) with a changed overall noise size can be computed using only the final SNR from the original evaluation. We used this trick to replace the two parameters N_F and N_C with the single parameter $\frac{N_F}{N_C}$.

We then used a grid search to optimize parameters for each presentation time. In this grid search we used the parameters listed in Table 3. These parameter values cover the range for p, q & C densely. For σ_θ we chose $\{0, \frac{3}{8}, \frac{7}{8}, \frac{12}{8}, \infty\} \times \sigma_\theta$—the orientation bandwidth of the filter—covering the range of qualitative behaviours for this parameter.[1] Similarly we set the linear noise factor N_F to $\{0, 0.1, 1, 10\} \times N_C$. By saving the likelihood value for each image combination separately we could extract this cube for different parts of the data.

The results of the grid search are displayed in figure 40, displaying the maximum likelihood found in the slice which sets the given parameter to the plotted value. Different lines give the values for the different $\frac{N_F}{N_C}$ values. The different panels are based on different subsets of the data.

1 0 represents normalization exclusively by channels with the same orientation and ∞ represents equal weighting of all orientations.

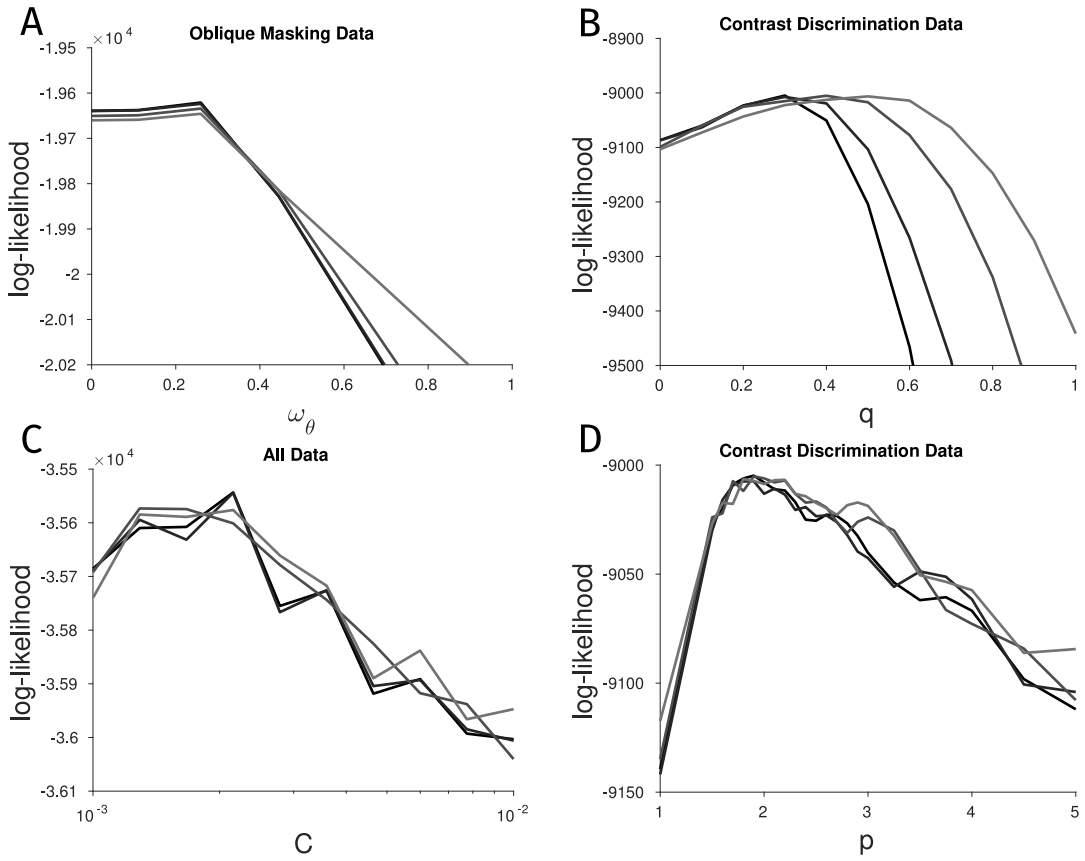

Figure 40: Evaluation of the grid search over the parameter space for the 1497ms presentation time. Each panel shows the maximum likelihood reached with the given parameter value. **A**: Bandwidth of the normalization pool evaluated over the oblique masking data. Maximum is at $\frac{7}{8}\sigma_\theta$. $\omega_\theta = \infty$ is plotted at $\omega_\theta = 1$. **B**: Exponent q evaluated on the contrast discrimination data. **C**: Constant C evaluated over all data. **D**: Exponent p evaluated on the contrast discrimination data.

Based on the displayed results on the grid search we drew the following conclusions:

- As displayed in Panel A, the clearly best bandwidth of the normalization pool ω_θ is the one slightly smaller than the bandwidth of the filter. ($\omega_\theta = \frac{7}{8}\sigma_\theta$)
- From Panel B: The composition of the noise and q are coupled. When the linear contribution to the noise grows, larger values of q are needed to compensate this. However, any $\frac{N_F}{N_C}$-ratio explains the data equally well, when we use the adequate q. To remove this ambiguity, we set N_F to zero.
- Finally, from Panel C+D: p and C are reasonably well constrained by the data. However, the oblique masking data and the contrast discrimination favour slightly different values for p and C (not shown). These result in the two parameter values we display in the main paper. The two parameters differ only slightly however and make reasonably similar predictions as we saw in the main paper.

We evaluated the same range of parameter values for the other presentation times and for single pixel normalization. For the other presentation times we can draw the same conclusions as above. For the single pixel normalization, however, we find a pronounced inconsistency of oblique masking and contrast discrimination. The oblique masking requires a much higher p value than the discrimination data. Consequently a parameter which optimizes the results for both conditions is considerably worse in the single pixel normalization model than in the mean normalization model.

As our grid was a bit coarse we used the best parameter from the grid search as a starting point for some further optimization with a BFGS algorithm employing the gradients from Appendix A.2:

- First, we fit the bandwidth ω_θ to the oblique masking data, fixing N_F to 0 and p, q & C to the optimal values from the grid.
- Using the estimate for ω_θ from this optimization, we fitted 4 parameter parameter values for p, q, C and N_C:
  - One for each presentation time to the corresponding contrast discrimination data, starting the optimization at the optimal value from the grid search for that presentation time.
  - One to the oblique masking data for the 1497ms presentation time, starting at the best grid point again.
  - One for the ModelFest dataset starting at the optimal parameter for the 1497ms presentation time from the gridsearch.

For an additional comparison on the ModelFest data, we fitted the ModelFest data adjusting only N_C starting from the parameter for $1497ms$ and $79ms$ respectively. To fit these we again calculated performance from the signal to noise ratios, reducing the computational cost for this step.

Finally, we fitted a parameter set for the ModelFest dataset specifically, although the estimates from the classical data were decent already. As we did not have individual percent correct values for these data, we transformed the given thresholds to surrogate blocks of trials with percent correct. We assumed 3 blocks of 100 trials each: One with 86 correct trials at the threshold, one with 100 correct trials at 1.5 times the threshold and a block with 50 correct trials a factor 3 below threshold. Using this surrogate data we

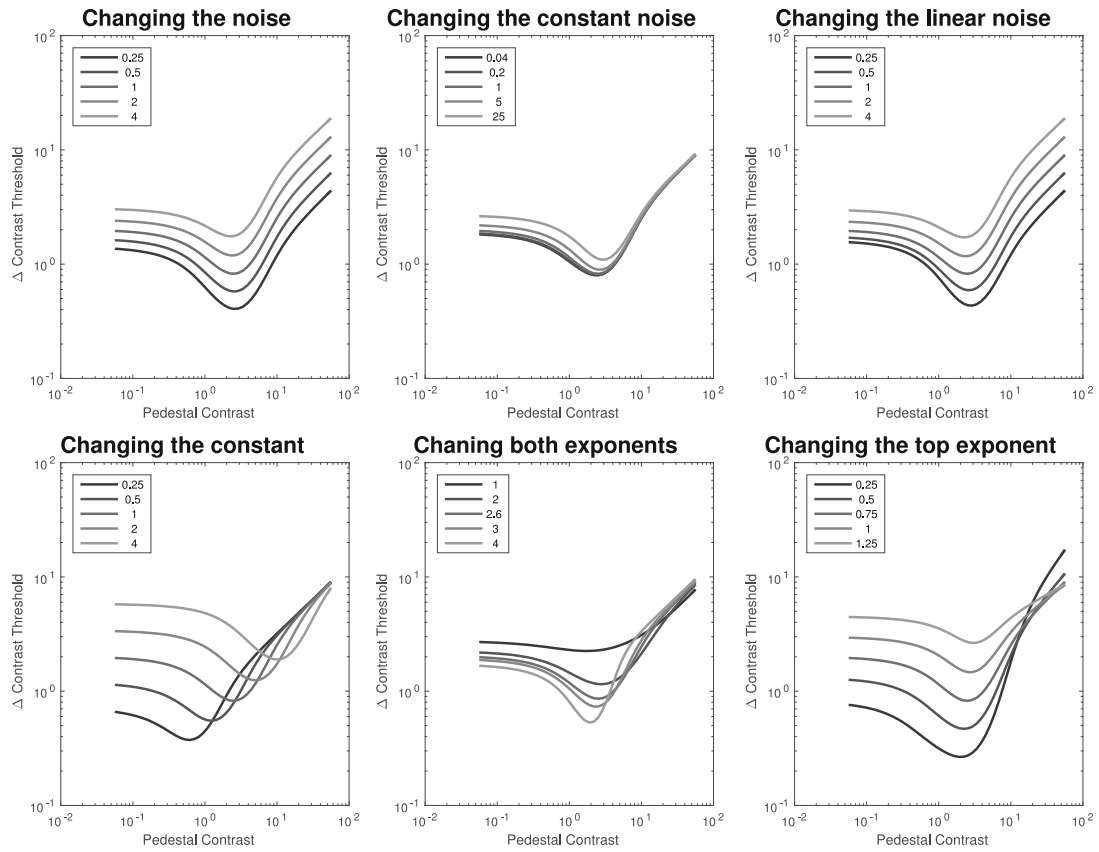

Figure 41: Effects of changing the parameters on contrast discrimination curves. Each panel shows the curve of contrast discrimination thresholds ("the dipper") when a single parameter of the model is changed leaving the others at their fitted values.

then fitted the normalization and noise parameters (N_c, N_f, C, p & q) as for the classical data.

As a last rather cosmetic step we refit the neuronal filter we employ with the final parameters to fit the data for detection well, which was necessary as the processing of the model does distort the csf (higher exponents exaggerate the differences between different input strengths).

To give the reader a better intuition, what the different parameter values mean, figure 41 shows the effect on the contrast discrimination results, when the different parameters are varied separately. Clearly the parameters N_f, N_c and C merely move the function around hardly changing its shape. In contrast changing p—i.e. both exponents—controls how peaked the dip in the contrast sensitivity function is. Changing q—i.e. only the numerator exponent—strongly affects detection performance and how steeply the discrimination thresholds rise with pedestal contrast for high pedestal contrasts.

A.2 DERIVATIVES OF THE MODEL

For parameter optimization we derived a gradient of the model likelihood with respect to all parameters. To compute this we also compute derivatives for the signal to noise ratio, percent correct and quite a few of the internal model states against each other. For a mathematically proficient reader these might thus provide some insight into the internal

dependencies of the model. Also these calculations illustrate that the derivatives of the stages in our model can be computed as for the now popular deep neural network models.

Our presentation of the model derivatives follows the calculation in backward order in analogy to the backprop algorithms for deep neural networks, i.e. we start with the likelihood and go back to the parameters using the chain rule consecutively. Computation can be implemented in forward order with equal ease.

For each step we will first calculate the derivatives with respect to the parameters used in the step directly and then the one to the input of the processing step, which allows the calculation of derivatives with respect to parameters used in the previous processing step.

A.2.1 *Likelihood from signal to noise ratio*

We start with the log-likelihood, which depends on the lapse rate λ and the signal to noise ratio $\frac{d}{\sqrt{\eta}}$ from the model. As a first step we calculate the derivative with respect to p_c the percent correct predicted by the model without lapses:

$$\frac{\partial l(correct)}{\partial p'_c} = \frac{\partial}{\partial p'_c} \log(\lambda + (1-2\lambda)p'_c) = \frac{1-2\lambda}{l + (1-2\lambda)p'_c} \tag{38}$$

$$\frac{\partial l(incorrect)}{\partial p'_c} = \frac{\partial}{\partial p'_c} \log(1-\lambda-(1-2\lambda)p'_c) = \frac{-(1-2\lambda)}{1-\lambda-(1-2\lambda)p'_c} \tag{39}$$

The derivative of the predicted percent correct p_c with regard to the signal to noise ratio $\frac{d}{\sqrt{\eta}}$ is simply the density of the normal distribution at the signal to noise ratio:

$$\frac{\partial p'_c}{\partial \left(\frac{d}{\sqrt{\eta}}\right)} = \phi\left(\frac{d}{\sqrt{\eta}}\right) \tag{40}$$

A.2.2 *Decoding*

Next we analyse the decoding stage. This stage receives 2 arrays of model responses $\{r_i^{(1)}\}$ and $\{r_i^{(2)}\}$ both indexed with an index i from an index-set $\mathcal{I}$ over position, orientation and frequency. As the output we consider the signal to noise ratio $\frac{d}{\sqrt{\eta}}$. To calculate the derivative of the signal to noise ratio $\frac{d}{\sqrt{\eta}}$ with respect to any parameter used earlier in the model x we use the following formulas:

$$\frac{\partial}{\partial x}\frac{d}{\sqrt{\eta}} = \frac{1}{\sqrt{\eta}}\frac{\partial d}{\partial x} + d\eta^{\frac{-3}{2}}\frac{\partial \eta}{\partial x} = \frac{1}{\sqrt{\eta}}\sum_{i\in\mathcal{I}}\frac{\partial d_i}{\partial x} + d\eta^{\frac{-3}{2}}\sum_{i\in\mathcal{I}}\frac{\partial \eta_i}{\partial x} \tag{41}$$

The two parameters of the decoding stage are the size of the constant noise N_c and the factor for the noise variance N_f for which we calculate the derivatives first:

$$\frac{\partial \eta_i}{\partial N_c} = 1 \qquad \frac{\partial \eta_i}{\partial N_F} = r_i \qquad \frac{\partial r_i}{\partial N_C} = \frac{\partial r_i}{\partial N_F} = 0 \tag{42}$$

For any other parameters x which changes r_i, we can calculate the derivatives from the derivative $\frac{\partial r_i}{\partial x}$:

$$\forall i \in \mathcal{I}: \frac{\partial \eta_i}{\partial x} = \frac{\partial}{\partial x}\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{n_i^{(1)} + n_i^{(2)}}(n_i^{(1)} + n_i^{(2)}) = \frac{\partial}{\partial x}\left(r_i^{(1)} - r_i^{(2)}\right)^2 \tag{43}$$

$$= 2\left(r_i^{(1)} - r_i^{(2)}\right)\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) \tag{44}$$

$$\frac{\partial d_i}{\partial x} = \frac{\partial}{\partial x}\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{\sqrt{n_i^{(1)} + n_i^{(2)}}} \tag{45}$$

$$= 2\frac{r_i^{(1)} - r_i^{(2)}}{\sqrt{n_i^{(1)} + n_i^{(2)}}}\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) - \frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{2\left(n_i^{(1)} + n_i^{(2)}\right)^{\frac{3}{2}}}\left(\frac{\partial n_i^{(1)}}{\partial x} + \frac{\partial n_i^{(2)}}{\partial x}\right) \tag{46}$$

$$= 2\frac{r_i^{(1)} - r_i^{(2)}}{\sqrt{n_i^{(1)} + n_i^{(2)}}}\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) - N_f\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{2\left(n_i^{(1)} + n_i^{(2)}\right)^{\frac{3}{2}}}\left(\frac{\partial r_i^{(1)}}{\partial x} + \frac{\partial r_i^{(2)}}{\partial x}\right) \tag{47}$$

using in the last step:

$$\frac{\partial n_i}{\partial x} = N_f\frac{\partial r_i}{\partial x} \tag{48}$$

A.2.3 *Normalization*

Thus in the Normalization stage we require the derivatives of the response r_i, which we again first compute for the parameters of this stage p, q, C and then for the bandwidths of the normalization ω and the filter σ.

for $a_i = 0$ all derivatives are 0 because r_i is then 0 independent of all parameters for $a_i > 0$:

$$\frac{\partial r_i}{\partial p} = \log(a_i)r_i - \frac{a_i^{p+q}}{\left(C^p + b_i\right)^2}\left[\log(C) + \sum_{i\in\mathcal{I}}(G * \log(a))(x_i)\right] \tag{49}$$

$$\frac{\partial r_i}{\partial q} = \log(a_i)r_i \tag{50}$$

$$\frac{\partial r_i}{\partial C} = -r_i p\frac{C^{p-1}}{C^p + b_i} \tag{51}$$

For computing the derivatives with respect to the σs we need to compute the derivatives of r_i towards a_i and b_i as well as the derivatives of the filter values:

$$\frac{\partial r_i}{\partial a_i} = (p+q)\frac{a_i^{p+q-1}}{C^p + b_i} + \frac{\partial r_i}{\partial b_i}\frac{\partial b_i}{\partial a_i} \tag{52}$$

$$\frac{\partial r_i}{\partial b_i} = -r_i\frac{1}{C^p + b_i} \tag{53}$$

A Gaussian $G(x|\sigma)$ in x without normalization (as the ones in frequency space to define the log-Gabors) has the following derivative with respect to its standard deviation σ

$$\frac{\partial G}{\partial \sigma}(x) = \frac{(x-\overline{x})^2}{\sigma^3}\exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right) \tag{54}$$

For the normalization of a Gaussians $G_n = G_n(\cdot|\sigma)$ we used the sum over the grid in each case, i.e.:

$$\forall i \in \mathcal{I}: \qquad G_n(x_i) = \frac{G(x_i)}{\sum_{j\in\mathcal{I}} G(x_j)} \tag{55}$$

The derivative of this normalized Gaussians G_n with respect to its standard deviation σ is thus given by:

$$\begin{aligned} \frac{\partial G_n}{\partial \sigma}(x_i) &= \frac{1}{\sum_{j\in\mathcal{I}} G(x_j)} \frac{\partial G}{\partial \sigma}(x_i) - \frac{G(x_i)}{(\sum_{j\in\mathcal{I}} G(x_j))^2} \sum_{j\in\mathcal{I}} \frac{\partial G}{\partial \sigma}(x_j) && (56) \\ &= \frac{1}{\sum_{j\in\mathcal{I}} G(x_j)} \left(\frac{\partial G}{\partial \sigma}(x_i) - G_n(x_i) \sum_{j\in\mathcal{I}} \frac{\partial G}{\partial \sigma}(x_j) \right) && (57) \end{aligned}$$

The same is true for a convolution, when only one of the two convolved functions f depends on the variable x:

$$\frac{\partial (f(x) * g(y))}{\partial x} = \frac{\partial f(x)}{\partial x} * g(y) \tag{58}$$

$$\frac{\partial (g(y) * f(x))}{\partial x} = g(y) * \frac{\partial f(x)}{\partial x} \tag{59}$$

Thus we can compute the derivatives of $B = \{b_i\}_{i\in I}$ interpreted as the four dimensional array of normalization inputs for the each channel at each position: For any of the standard deviations $\omega_{x,y,\phi,f}$ we can decompose the 4D Gaussian into the 4 one dimensional convolutions and compute the 4 derivatives as follows:

$$\frac{\partial B}{\partial \omega_x} = \frac{\partial G(\omega_x)}{\partial \omega_x} * G(\omega_y, \omega_\phi, \omega_f) * A^p \tag{60}$$

$$\frac{\partial B}{\partial \omega_y} = \frac{\partial G(\omega_y)}{\partial \omega_y} * G(\omega_x, \omega_\phi, \omega_f) * A^p \tag{61}$$

$$\frac{\partial B}{\partial \omega_\phi} = \frac{\partial G(\omega_\phi)}{\partial \omega_\phi} * G(\omega_x, \omega_y, \omega_f) * A^p \tag{62}$$

$$\frac{\partial B}{\partial \omega_f} = \frac{\partial G(\omega_f)}{\partial \omega_f} * G(\omega_x, \omega_y, \omega_\phi) * A^p \tag{63}$$

For any parameter, except the parameters of the normalization pool:

$$\begin{aligned} \frac{\partial B}{\partial A} &= G(\omega_x, \omega_y, \omega_\phi, \omega_f) * \frac{\partial A^p}{\partial A} && (64) \\ &= G(\omega_x, \omega_y, \omega_\phi, \omega_f) * pA^{p-1} && (65) \end{aligned}$$

A.2.4 *Decomposition*

As we did not fit the filters to data we do not require the derivatives to their bandwidths for fitting. These derivatives can be computed nonetheless as follows:

The derivative of the absolute value we apply between the decomposition and the nonlinearity with respect to a parameter which influences real $\mathfrak{R}$ and imaginary $\mathfrak{I}$ part of a complex number z is:

$$\frac{\partial |f(x)|}{\partial x} = \frac{\mathfrak{R}(f(x))}{|f(x)|} \frac{\partial \mathfrak{R}(f(x))}{\partial x} + \frac{\mathfrak{I}(f(x))}{|f(x)|} \frac{\partial \mathfrak{I}(f(x))}{\partial x} \tag{66}$$

This is not a proper complex derivative, but only a real derivative by interpreting the complex z as $\in \mathbb{R}^2$.

Finally, to compute the derivatives of the filter output against the filter parameters, we can use the following formula

$$\frac{\partial \mathcal{F}(f)}{\partial x} = \mathcal{F}\left(\frac{\partial f}{\partial x}\right), \tag{67}$$

because the Fourier-transform is a linear operator.

The filtering in Fourier space is an element wise multiplication. Thus the derivative of a channel response $f(x, y)$ can be computed from the derivatives of the filters in Fourier space $g(x, y)$ and the image $I(x, y)$:

$$\frac{\partial f(x)}{\partial \sigma} = \frac{\partial}{\partial \sigma}\mathcal{F}^{-1}\left(\mathcal{F}(I(x,y))g(x,y)\right) = \mathcal{F}^{-1}\left(\mathcal{F}(I(x,y))\frac{\partial g(x,y)}{\partial \sigma}\right) \tag{68}$$

A.2.5 *Preprocessing*

Our preprocessing is an affine transformation. Thus the derivatives with respect to the original inputs can be computed from derivatives with respect to the preprocessed image simply by applying the same filters with flipped phase and adding back the mean luminance.

A.3 OPTIMIZING STIMULI

One analysis to compare different models, one interesting method is to optimize stimuli which are especially different or similar according to the model while keeping similarity according to another model constant (Wang & Simoncelli, 2008). As analyses of this type are a strength of image-computable models we use it in the main text to show the advantages of having an image-computable model. In this appendix we explain the details of the optimization procedure we employ to get the stimuli.

We aim to find luminance images ($I_1, I_2 \in \mathbb{R}^{N\times N}$) which have a given Root Mean Square Error ($RMSE_0$) from a given start image I_0 after conversion to contrast and cut out of the fovea and are either maximally easy (I_1)or maximally hard (I_2) to differentiate from I_0 according to the model.

Furthermore we require two constraints on the images to yield displayable and interesting stimuli: (i) All pixels must be in the range $[0, L_m]$ for a maximal displayable luminance L_m. (ii) Pixels for which the foveal window $w \in \mathbb{R}^{N\times N}$ is 0 shall be equal to the corresponding pixels in I_0.

For notation we shall use:

- N for the size of the square images
- $I' = w \cdot \left(I/(\frac{1}{N^2}\sum_{j,k=1}^{N} I_{0jk})\right)$ for the image I after conversion to contrast and application of the foveal window w. Here / means element-wise division. Note that we always use the mean luminance of I_0 for this conversion.
- $RMSE(I'_1, I'_0) = \sqrt{MSE(I'_1, I'_0)} = \sqrt{\sum_{j,k=1}^{N}(I'_{1jk} - I'_{0jk})^2}$ to denote the root mean squared error of two (converted) images I_0 and I_1.
- $d'(I_1, I_0)$ to denote the discriminability d' of I_1 and I_0 according to our model. Additionally we write $d'(I'_1, I'_0) := d'(I_1, I_0)$ for converted images overloading notation.

To allow a conversion back from a contrast image I' to a luminance image I we set (suppressing indices):

$$I(I') = \begin{cases} I_0 & w \leq 0.001 \\ \frac{1}{N^2} \sum_{j,k=1}^{N} I_{0jk})(I'/w) & w > 0.001 \end{cases} \tag{69}$$

, i.e. wherever the foveal window is 0 we set the luminance image to be equal to I_0. As we enforce I_1 and I_2 to be equal to I_0 there this yields correct results. To avoid numerical issues with the division by w we extend this enforced equality to pixels with $w \leq 0.001$

This yields the following optimization problem in mathematical short-hand:

Minimize $d'(I_1, I_0)$ / Maximize $d'(I_2, I_0)$

subject to:

$$0 \leq I_1, I_2 \leq L_m \tag{70}$$

$$RMSE(I'_2, I'_0) = RMSE(I'_1, I'_0) = RMSE_0 \tag{71}$$

$$\forall j, k = 1 \ldots N : w_{jk} \leq 0.001 \Rightarrow I_{1jk} = I_{2jk} = I_{0jk} \tag{72}$$

To solve this problem with nonlinear equality constraints approximately we relax the constraints quadratically and add one common parameter β which shall increase during optimization to increase the penalty for missing the constraints. Finally we add another regularizer $\sum_{j,k=1}^{N}(1 - w_{jk})^{\beta}(I'_{0jk} - I'_{1jk})^2$ which pushes the optimization to yield similar images near the edges of the window w.

For I_1 this yields the following relaxed optimization problem:

Minimize:

$$d'(I_1, I_0) + \beta^4 (MSE(I'_1, I'_0) - RMSE_0^2)^2 \tag{73}$$

$$+ \beta^2 \sum_{j,k=1}^{N} (1 - w_{jk})^{\beta}(I'_{0jk} - I'_{1jk})^2 \tag{74}$$

subject to:

$$0 < I_1 < L_m \tag{75}$$

We then use a gradient decent algorithm to solve this optimization problem starting from Gaussian noise added to the area where $w > .001$ with the correct $RMSE_0$ and cut to fit the displayable luminance range. We then apply a gradient decent during which we test at each point whether it is better than the previous one. Depending on the outcome of this we adjust the stepsize adding 30% every time we update successfully and dividing

by 2 every time we fail. We increase β by 1 every time the change predicted by the current gradient and step size is smaller than 0.001. When $\beta = 100$ and the predicted change is smaller than 10^{-6} we end the optimization. If at any time a pixel leaves the allowed luminance range we set it back inside the range by the smallest possible numerical value.

Aagten-Murphy, D., & Bays, P. M. (2017). Automatic and intentional influences on saccade landing. *Journal of Neurophysiology*, *118*(2), 1105–1122.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* (Software available from tensorflow.org)

Acerbi, L., & Ma, W. J. (2017). Practical bayesian optimization for model fitting with bayesian adaptive direct search. In *Advances in neural information processing systems* (pp. 1834–1844).

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, *37*(6), 1453–1467.

Adler, W. T., & Ma, W. J. (2017). Human confidence reports account for sensory uncertainty but in a non-bayesian way. *bioRxiv*, 093203.

Akaike, H. (1974). A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers Transactions on Automatic Control*, *19*(6), 716–723.

Alam, M. M., Patil, P., Hagan, M. T., & Chandler, D. M. (2015). A computational model for predicting local distortion visibility via convolutional neural network trained on natural scenes. In *2015 IEEE Internationa Conference on image Processing (ICIP)* (pp. 3967–3971). IEEE.

Alam, M. M., Vilankar, K. P., Field, D. J., & Chandler, D. M. (2014). Local masking in natural images: A database and analysis. *Journal of Vision*, *14*(8), 22:1–38.

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin & Review*, *23*(6), 1794–1801.

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 9:1–22.

Andrieu, C., & Roberts, G. (2009). The pseudo-marginal approach for efficient Monte-Carlo computations. *The Annals of Statistics*, *37*, 697–725.

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, *18*(4), 343–373.

Ash, R. B. (1990). *Information theory.* Dover Publications Inc., New York.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183-193.

Açık, 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), 1541–1553.

Badcock, D. R. (1984). Spatial phase or luminance profile discrimination? *Vision Research*, *24*(6), 613–623.

Badcock, D. R. (1988). Discrimination of spatial phase changes: Contrast and position codes. *Spatial Vision*, *3*(4), 305–322.

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: Methodology and applications with R.* London: Chapman and Hall/CRC Press.

Baddeley, A., & Turner, R. (2005). spatstat: An R package for analyzing spatial point

patterns. *Journal of Statistical Software*, *12*(6), 1–42.

Baddeley, A., Turner, R., Mateu, J., & Bevan, A. (2013). Hybrids of Gibbs point process models and their implementation. *Journal of Statistical Software*, *55*(11), 1–43.

Baker, D. H., Meese, T. S., & Georgeson, M. A. (2007). Binocular interaction: Contrast matching and contrast discrimination are predicted by the same model. *Spatial Vision*, *20*(5), 397–413.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13:1–18.

Baldwin, A. S., Meese, T. S., & Baker, D. H. (2012). The attenuation surface for contrast sensitivity has the form of a witch's hat within the central visual field. *Journal of Vision*, *12*(11), 23:1–17.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*(3), 241–253.

Barlow, H. B. (1969). Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences*, *156*(2), 872–881.

Barthelmé, S., Trukenbrod, H., Engbert, R., & Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, *13*(12), 1:1–34.

Barthelmé, S., & Chopin, N. (2011). ABC-EP: Expectation propagation for likelihoodfree bayesian computation. In *International conference on mashine learning* (pp. 289–296).

Barthelmé, S., & Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, *109*(505), 315–333.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beaumont, M. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, *164*, 1139–1160.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., ... Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Bedford, R. E., & Wyszecki, G. (1957). Axial chromatic aberration of the human eye. *Journal of the Optical Society of America*, *47*(6), 564–565.

Berger, J. O. (2013). *Statistical decision theory and bayesian analysis*. Springer Science & Business Media.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: Ideas and concepts*. San Francisco, USA: Holden-Day.

Bird, C. M., Henning, G. B., & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A*, *19*(7), 1267–1273.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*(298), 269–306.

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, *203*(1), 237–260.

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. (2014). What/where to look next? Modeling top-

down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(5), 523–538.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010* (pp. 177–186). Springer.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Bradley, C., Abrams, J., & Geisler, W. S. (2014). Retina-V1 model of detectability across the visual field. *Journal of Vision*, *14*(12), 22:1–22.

Brainard, D. H. (2015). Color and the Cone Mosaic. *Annual Review of Vision Science*, *1*(1), 519–546.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art.* Univ. Chicago Press.

Buzsáki, G., & Mizuseki, K. (2014). The log-dynamic brain: How skewed distributions affect network operations. *Nature Reviews Neuroscience*, *15*(4), 264–278.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2016). *MIT saliency benchmark.* http://saliency.mit.edu/.

Cadieu, C. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Cajar, A., Engbert, R., & Laubrock, J. (2016). Spatial frequency processing in the central and peripheral visual field during scene viewing. *Vision Research*, *127*, 186–197.

Cajar, A., Schneeweiß, P., Engbert, R., & Laubrock, J. (2016). Coupling of attention and saccades when viewing scenes with central and peripheral degradation. *Journal of Vision*, *16*(2), 8:1-19.

Campbell, F. W., Cleland, B. G., Cooper, G. F., & Enroth-Cugell, C. (1968). The angular selectivity of visual cortical cells to moving gratings. *The Journal of Physiology*, *198*(1), 237–250.

Campbell, F. W., & Kulikowski, J. J. (1966). Orientational selectivity of the human visual system. *The Journal of Physiology*, *187*(2), 437–445.

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, *197*(3), 551–566.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Carandini, M., & Ringach, D. L. (1997). Predictions of a recurrent model of orientation selectivity. *Vision Research*, *37*(21), 3061–3071.

Carello, C. D., & Krauzlis, R. J. (2004). Manipulating intent: Evidence for a causal role of the superior colliculus in target selection. *Neuron*, *43*(4), 575–583.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of*

Statistical Software, *76*(1).

Carpenter, R. (1999). A neural mechanism that randomises behaviour. *Journal of Consciousness Studies*, *6*(1), 13–22.

Carter, B. E., & Henning, G. B. (1971). The detection of gratings in narrow-band visual noise. *Journal of Physiology*, *219*(2), 355-365.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, *46*(3), 167–174.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology*, *62*(1), 1–14.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6:1–15.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2530–2546.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2547–2556.

Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision*, *16*(10), 13:1–18.

Chapman, P., Underwood, G., & Roberts, K. (2002). Visual search patterns in trained and untrained novice drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 157–167.

Charman, W. N., & Jennings, J. A. M. (1976). Objective measurements of the longitudinal chromatic aberration of the human eye. *Vision Research*, *16*(9), 999–1005.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*(24), 4118–4133.

Chopin, N., Jacob, P., & Papaspiliopoulos, O. (2013). SMS^2: An efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society B*, *75*, 397–426.

Clarke, A. D. F., Mahon, A., Irvine, A., & Hunt, A. R. (2017). People are unable to recognize or report on their own eye movements. *Quarterly Journal of Experimental Psychology*, *70*(11), 2251–2270.

Clarke, A. D. F., Stainer, M. J., Tatler, B. W., & Hunt, A. R. (2017). The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of Vision*, *17*(11), 12:1–19.

Clarke, A. D. F., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: A time-window-of-integration model. *Journal of Cognitive Neuroscience*, *16*(6), 1000–1009.

Conover, W. J. (1980). *Practical nonparametric statistics*. New York, USA: Wiley.

Cooper, E. A., Piazza, E. A., & Banks, M. S. (2012). The perceptual basis of common photographic practice. *Journal of Vision*, *12*(5), 8:1–14.

Cornelissen, T. H. W., & Võ, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics*, *79*(1), 154–168.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, UK: Cambridge University Press.

Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, *300*(1), 5–25.

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *The Journal of Comparative Neurology*, *292*(4), 497–523.

Curcio, C. A., Sloan, K. R., Packer, O., Hendrickson, A. E., & Kalina, R. E. (1987). Distribution of cones in human and monkey retina: Individual variability and radial asymmetry. *Science*, *236*(4801), 579–582.

Dacey, D. M., & Petersen, M. R. (1992). Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proceedings of the National Academy of sciences*, *89*(20), 9666–9670.

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, *20*(10), 847–856.

Derrington, A. M., & Henning, G. B. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, *29*(2), 241–246.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*(1373), 1245–1255.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545–559.

Diaz, J. A., Queirazza, F., & Philiastides, M. G. (2017). Perceptual learning alters post-sensory processing in human decision-making. *Nature Human Behaviour*, *1*(2), 35:1-9.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327.

Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)* (pp. 1–6).

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)* (pp. 1–7).

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 28:1-17.

Doucet, A., de Freitas, N., & (eds.), N. G. (2001). *Sequential Monte Carlo methods in practice*. Berlin Heidelberg New York: Springer-Verlag.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222.

Duncan, R. O., & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, *38*(4), 659–671.

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., & Nuthmann, A. (2016). Salient in space, salient in time: Fixation probability predicts fixation duration during natural scene viewing. *Journal of Vision*, *16*(11), 13:1-17.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of*

Vision, *8*(2), 2:1–19.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18:1-26.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, *24*(6), 581–604.

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

Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences*, *103*(18), 7192–7197.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Engbert, R., Rothkegel, L. O. M., Backhaus, D., & Trukenbrod, H. A. (2016). *Evaluation of velocity-based saccade detection in the SMI-ETG 2W system* (Tech. Rep.). Potsdam, Germany: Allgemeine und Biologische Psychologie, Universität Potsdam.

Engbert, R., Trukenbrod, H. A., Barthelme, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1), 14:1-17.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cerebral Cortex*, *7*(2), 181–192.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A*, *11*(6), 1710–1719.

Foley, J. M., & Boynton, G. M. (1994). New model of human luminance pattern vision mechanisms: analysis of the effects of pattern orientation, spatial phase, and temporal frequency. In *SPIE procedings: Computational Vision Based on Neurobiology* (Vol. 2054, pp. 32–42).

Foley, J. M., & Chen, C. (1997). Analysis of the effect of pattern adaptation on pattern pedestal effects: A two-process model. *Vision Research*, *37*(19), 2779–2788.

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, *21*(7), 1041–1053.

Foley, J. M., Varadharajan, S., Koh, C. C., & Farias, M. C. Q. (2007). Detection of gabor patterns of different sizes, shapes, phases and eccentricities. *Vision Research*, *47*(1), 85–107.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, *48*(17), 1777–1790.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recog-

nition. *Journal of Vision*, *8*(2), 6:1-17.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(9), 891–906.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495.

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., ... Tolias, A. S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature Neuroscience*, *17*(6), 851–857.

Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, *3*(6), 535–536.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Computation*, *26*(10), 2103–2134.

Gattass, R., Gross, C. G., & Sandell, J. H. (1981). Visual topography of V2 in the macaque. *The Journal of Comparative Neurology*, *201*(4), 519–539.

Gattass, R., Sousa, A. P., & Gross, C. G. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, *8*(6), 1831–1845.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *arXiv:1505.07376 [cs, q-bio]*.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*(7), 563–572.

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv:1706.06969 [cs, q-bio, stat]*.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781.

Geisler, W. S., & Albrecht, D. G. (1995). Bayesian analysis of identification performance in monkey visual cortex: Nonlinear mechanisms and stimulus certainty. *Vision Research*, *35*(19), 2723-2730.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Taylor & Francis.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gelman, A., Roberts, G., & Gilks, W. (1996). Efficient metropolis jumping hules. *Bayesian Statistics*, *5*(42), 599-608.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Georgeson, M. A., & Meese, T. S. (1997). Perception of stationary plaids: The role of spatial filters in edge analysis. *Vision Research*, *37*(23), 3255–3271.

Georgeson, M. A., & Meese, T. S. (2006). Fixed or variable noise in contrast discrimina-

tion? The jury's still out.... *Vision Research*, *46*(25), 4294–4303.

Georgeson, M. A., Wallis, S. A., Meese, T. S., & Baker, D. H. (2016). Contrast and lustre: A model that accounts for eleven different forms of contrast discrimination in binocular vision. *Vision Research*, *129*, 98–118.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Gneiting, T., Balabdaoui, F., & Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, *69*, 243–268.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley.

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865.

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. A. (2013). A neural population model for visual pattern detection. *Psychological Review*, *120*(3), 472–496.

Goris, R. L. T., Simoncelli, E. P., & Movshon, J. A. (2015). Origin and Function of Tuning Diversity in Macaque Visual Cortex. *Neuron*, *88*(4), 819–831.

Goris, R. L. T., Zaenen, P., & Wagemans, J. (2008). Some observations on contrast detection in noise. *Journal of Vision*, *8*(9), 4:1–15.

Graham, N. (1989). *Visual Pattern Analyzers.* Oxford University Press.

Graham, N., & Nachmias, J. (1971). Detection of grating patterns containing two spatial frequencies: A comparison of single-channel and multiple-channels models. *Vision Research*, *11*(3), 251–259.

Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, *62*, 1–8.

Haario, H., Laine, M., Mira, A., & Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, *16*(4), 339–354.

Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, *7*(2), 223–242.

Hahn, L. W., & Geisler, W. S. (1995). Adaptation mechanisms in spatial vision-I. Bleaches and backgrounds. *Vision Research*, *35*(11), 1585–1594.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*(5), 347–356.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, *18*(10), 1279–1296.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Neural information prcessing systems* (Vol. 1, p. 5).

Harvey, B. M., & Dumoulin, S. O. (2011). The relationship between cortical magnification factor and population receptive field size in human visual cortex: Constancies in cortical architecture. *Journal of Neuroscience*, *31*(38), 13604–13612.

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*(1), 59–67.

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, *17*(5), 23:1–

17.

Hays, W. L. (1994). *Statistics*. Wadsworth Publishing, Independence, KY.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences*, *93*(2), 623–627.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. V. Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements* (pp. 537–562). Oxford: Elsevier.

Henderson, J. M., Weeks Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228.

Hennig, P., & Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Mashine Learning Research*, *13*(1), 1809–1837.

Henning, G. B., Hertz, B. G., & Broadbent, D. (1975). Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision Research*, *15*(8–9), 887–897.

Hernández-Lobato, J. M., Hoffman, M. W., & Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems* (pp. 918–926).

Hesse, H. (1922). *Siddharta. Eine indische Dichtung*. Berlin, Germany: Fischer.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence* (Vol. 8). Oxford, England: U Michigan Press.

Holmes, D. J., & Meese, T. S. (2004). Grating and plaid masks indicate linear summation in a contrast gain pool. *Journal of Vision*, *4*(12), 1080-1089.

Hood, D. C. (1998). Lower-level visual processing and models of light adaptation. *Annual Review of Psychology*, *49*(1), 503–535.

Hood, D. C., & Finkelstein, M. (1986). Sensitivity to light. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance (Vol. 1: Sensory processes and perception)*. John Wiley and Sons, New York.

Hooge, I. T. C., & Erkelens, C. J. (1998). Adjustment of fixation duration in visual search. *Vision Research*, *38*(9), 1295–1302.

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), 1901–1908.

Houck, C. R., Joines, J., & Kay, M. G. (1995). *A genetic algorithm for function optimization: A Matlab implementation* (Vol. 95; Tech. Rep. No. 09). Raleigh, NC, USA: North Carolina State University-IE Transactions.

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 262–270).

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Hughes, A. E., Southwell, R. V., Gilchrist, I. D., & Tolhurst, D. J. (2016). Quantifying peripheral and foveal perceived differences in natural image patches to predict visual search performance. *Journal of Vision*, *16*(10), 18:1–17.

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Hulleman, J., & Olivers, C. N. L. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*, e132.

Illian, J. B., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons.

Illian, J. B., Sørbye, S. H., & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *The Annals of Applied Statistics*, *6*(4), 1499–1530.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10), 1489–1506.

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

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: Toward a unifying model. *Journal of the Optical Society of America A*, *17*(11), 1899–1917.

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

Jackson, E. A. (1992). *Perspectives of nonlinear dynamics*. Cambridge University Press.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218).

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *186*(1007), 453–461.

Jennings, J. A. M., & Charman, W. N. (1981). Off-axis image quality in the human eye. *Vision Research*, *21*(4), 445–455.

Johnston, K., & Everling, S. (2011). Frontal cortex and flexible control of saccades. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook on eye movements* (pp. 279–302).

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*(4), 455–492.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics*, *43*(4), 346–354.

Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations* (Tech. Rep.). Massachusetts institute of technology, cambridge , MA 02139 USA: MIT computer science and artificial intelligence laboratory.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th international conference on computer vision* (pp. 2106–2113). IEEE.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is

the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 1307-1322.

Kelly, D. H. (1979). Motion and vision II: Stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America*, *69*(10), 665-672.

Keshvari, S., & Rosenholtz, R. (2016). Pooling of continuous features provides a unifying account of crowding. *Journal of Vision*, *16*(3), 39:1–15.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), 1–29.

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), 7:1–15.

Kingdom, F. A. (2016). Fixed versus variable internal noise in contrast transduction: The significance of Whittle's data. *Vision Research*, *128*, 1–5.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, *34*(5-6), 975–986.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680.

Klein, C., & Foerster, F. (2001). Development of prosaccade and antisaccade task performance in participants aged 6 to 26 years. *Psychophysiology*, *38*(2), 179–189.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, *10*(4), 346–352.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 238:1–12.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Koenderink, J., Valsecchi, M., Doorn, A. v., Wagemans, J., & Gegenfurtner, K. (2017). Eidolons: Novel stimuli for vision research. *Journal of Vision*, *17*(2), 7:1–36.

Kok, E. M., Aizenman, A. M., Võ, M. L.-H., & Wolfe, J. M. (2017). Even if I showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, *17*(12), 2:1–11.

Kontsevich, L. L., Chen, C.-C., & Tyler, C. W. (2002). Separating the effects of response nonlinearity and internal noise psychophysically. *Vision Research*, *42*(14), 1771–1784.

Korattikara, A., Chen, Y., & Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International conference on machine learning* (pp. 181–189).

Kortum, P. T., & Geisler, W. S. (1995). Adaptation mechanisms in spatial vision ii: Flash thresholds and background adaptation. *Vision Research*, *35*(11), 1595–1609.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417-446.

Kruthiventi, S. S. S., Ayush, K., & Babu, R. V. (2015). DeepFix: A fully convolutional neural network for predicting human eye fixations. *arXiv:1510.02927 [cs]*.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2017). Saliency benchmarking: Separating models, maps and metrics. *arXiv preprint arXiv:1704.08615*.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563 [cs, q-bio, stat]*.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*(6483), 742–744.

Land, M. F., 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.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Laparra, V., Ballé, J., Berardino, A., & Simoncelli, E. P. (2016). Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging*, *2016*(16), 1–6.

Laparra, V., Berardino, A., Ballé, J., & Simoncelli, E. P. (2017). Perceptually optimized image rendering. *arXiv:1701.06641 [cs]*.

Laubrock, J., Cajar, A., & Engbert, R. (2013). Control of fixation duration during scene viewing by interaction of foveal and peripheral processing. *Journal of Vision*, *13*(12), 11:1–20.

Laughlin, S. (1983). Matching coding to scenes to enhance efficiency. In O. J. Braddick & A. C. Sleigh (Eds.), *Physical and biological processing of images* (pp. 42–52). Springer.

Law, K., Stuart, A., & Zygalakis, K. (2015). *Data assimilation*. New York: Springer-Verlag.

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436-444.

Legge, G. E. (1984a). Binocular contrast summation—I. Detection and discrimination. *Vision Research*, *24*(4), 373–383.

Legge, G. E. (1984b). Binocular contrast summation—II. Quadratic summation. *Vision Research*, *24*(4), 385–394.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America*, *70*(12), 1458–1471.

Legge, G. E., Kersten, D., & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A*, *4*(2), 391–404.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, *45*(1), 251–266.

Le Meur, O., & Coutrot, A. (2016). Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Research*, *121*, 72–84.

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*, *116*, 152–164.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, *90*(1), 204–217.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in cognitive sciences*, *6*(1), 9–16.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P.

(2014). Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*.

Ludwig, C. J. H., & Gilchrist, I. D. (2002). Measuring saccade curvature: A curve-fitting approach. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 618–624.

Ma, W. J., Shen, S., Dziugaite, G., & van den Berg, R. (2015). Requiem for the max rule? *Vision Research*, *116*, 179–193.

Marin, J.-M., & Robert, C. (2007). *Bayesian core: A practical approach to computational bayesian statistics*. Springer Science & Business Media.

May, K. A., & Solomon, J. A. (2015a). Connecting psychophysical performance to neuronal response properties I: Discrimination of suprathreshold stimuli. *Journal of Vision*, *15*(6), 8:1-26.

May, K. A., & Solomon, J. A. (2015b). Connecting psychophysical performance to neuronal response properties II: Contrast decoding and detection. *Journal of Vision*, *15*(6), 9:1–21.

May, K. A., & Zhaoping, L. (2016). Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns. *Current Biology*, *26*(12), 1571–1576.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. CRC press.

Mechler, F., Reich, D. S., & Victor, J. D. (2002). Detection and discrimination of relative spatial phase by V1 neurons. *Journal of Neuroscience*, *22*(14), 6129–6157.

Meese, T. S., Georgeson, M. A., & Baker, D. H. (2006). Binocular contrast vision at and above threshold. *Journal of Vision*, *6*(11), 1224–1243.

Meese, T. S., & Holmes, D. J. (2002). Adaptation and gain pool summation: Alternative models and masking data. *Vision Research*, *42*(9), 1113–1125.

Meng, X., & Qian, N. (2005). The oblique effect depends on perceived, rather than physical, orientation and direction. *Vision Research*, *45*(27), 3402–3413.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Mohr, J., Seyfarth, J., Lueschow, A., Weber, J. E., Wichmann, F. A., & Obermayer, K. (2016). BOiS—berlin object in scene database: Controlled photographic images for visual search experiments with quantified contextual priors. *Frontiers in Psychology*, *7*, 749:1–6.

Mokler, A., & Fischer, B. (1999). The recognition and correction of involuntary prosaccades in an antisaccade task. *Experimental Brain Research*, *125*(4), 511–516.

Munoz, D. P., & Everling, S. (2004). Look away: The anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience*, *5*(3), 218–228.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Müller, H. J., & Krummenacher, J. (2006). Visual search and selective attention. *Visual Cognition*, *14*(4-8), 389–410.

Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: Discrimination may be better than detection. *Vision Research*, *14*(10), 1039–1042.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3), 4:1-14.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation

selection in visual search. *Vision Research*, *49*(10), 1286–1294.

Naka, K. I., & Rushton, W. a. H. (1966). S-potentials from colour units in the retina of fish (Cyprinidae). *The Journal of Physiology*, *185*(3), 536–555.

Navarro, R., Williams, D. R., & Artal, P. (1993). Modulation transfer of the human eye as a function of retinal eccentricity. *Journal of the Optical Society of America A*, *10*(2), 201–212.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain monte carlo* (Vol. 2, pp. 113–162).

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621.

Nicholls, J. G., Martin, A. R., Fuchs, P. A., Brown, D. A., Diamond, M. E., & Weisblat, D. A. (2012). *From neuron to brain* (5th ed.). Sinauer Associates, Sunderland/MA.

Nowakowska, A., Clarke, A. D. F., & Hunt, A. R. (2017). Human visual search behaviour is far from ideal. *Proceedings of the Royal Society of London B: Biological Sciences*, *284*(1849), 20162767.

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review*, *24*(2), 370–392.

Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 82–96.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, *11*.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405.

Nyström, M., & Holmqvist, K. (2008). Semantic override of low-level features in image viewing–Both initially and overall. *Journal of Eye Movement Research*, *2*(2), 2:1–11.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology*, *5*(3), 377–391.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Over, E., Hooge, I., Vlaskamp, B., & Erkelens, C. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, *47*(17), 2272–2280.

Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i Nieto, X. (2017). SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. *arXiv:1701.01081 [cs]*.

Paré, M., & Dorris, M. C. (2011). The role of posterior parietal cortex in the regulation of saccadic eye movements. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook on eye movements* (pp. 257–278).

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, *2*(9), 1508–1532.

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across

face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314–E3323.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225.

Phillips, G. C., & Wilson, H. R. (1984). Orientation bandwidths of spatial mechanisms measured by masking. *Journal of the Optical Society of America A*, *1*(2), 226–232.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Pointer, J. S., & Hess, R. F. (1989). The contrast sensitivity gradient across the human visual field: With emphasis on the low spatial frequency range. *Vision Research*, *29*(9), 1133–1151.

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, *33*(7), 993–999.

Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J., Carli, M., & Battisti, F. (2008). Color image database for evaluation of image quality metrics. In *IEEE 10th Workshop on Multimedia Signal Processing* (pp. 403–408).

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), *Attention and performance* (Vol. 10, pp. 531–556). Hilldale, NJ, USA: Erlbaum.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(9), 970–982.

Ramos Gameiro, R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and exploitation in natural viewing behavior. *Scientific Reports*, *7*, 2311:1-27.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438-481.

Reich, S., & Cotter, C. (2015). *Probabilistic forecasting and Bayesian data assimilation.* Cambridge, UK: Cambridge University Press.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476.

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook on eye movements* (pp. 528–550).

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in

macaque primary visual cortex. *Journal of Neurophysiology*, *88*(1), 455–463.

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation selectivity in macaque V1: Diversity and laminar dependence. *The Journal of Neuroscience*, *22*(13), 5639–5651.

Robert, C. P., & Casella, G. (2009). *Introducing Monte Carlo methods with R*. New York: Springer.

Robert, C. P., & Casella, G. (2013). *Monte Carlo statistical methods*. Berlin: Springer.

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367.

Robson, J. G., & Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, *21*(3), 409–418.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, *73*(2), 713–726.

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: Reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A*, *22*(5), 801–809.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, *2*(1), 437–457.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4), 14:1–17.

Rothkegel, L. O. M., Schütt, H. H., Trukenbrod, H. A., Wichmann, F. A., & Engbert, R. (2018). Searchers adjust their eye movement dynamics to the target characteristics in natural scenes. *arXiv:1802.04069 [q-bio]*.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2016). Influence of initial fixation position in scene viewing. *Vision Research*, *129*, 33–49.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, *17*(13), 3:1-18.

Rovamo, J., Franssila, R., & Näsänen, R. (1992). Contrast sensitivity as a function of spatial frequency, viewing distance and eccentricity with and without spatial noise. *Vision Research*, *32*(4), 631–637.

Rovamo, J., Luntinen, O., & Näsänen, R. (1993). Modelling the dependence of contrast sensitivity on grating area and spatial frequency. *Vision Research*, *33*(18), 2773–2788.

Rovamo, J., Mustonen, J., & Näsänen, R. (1994). Modelling contrast sensitivity as a function of retinal illuminance and grating area. *Vision Research*, *34*(10), 1301–1314.

Rovamo, J., & Virsu, V. (1979). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, *37*(3), 495–510.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650.

Sammaknejad, N., Pouretemad, H., Eslahchi, C., Salahirad, A., & Alinejad, A. (2017). Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology*, *13*(3), 232–240.

Santella, A., & DeCarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 symposium on eye tracking research & applications* (pp. 27–34). New York, NY, USA: ACM.

Schomaker, J., Walper, D., Wittmann, B. C., & Einhäuser, W. (2017). Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience. *Vision Research*, *133*, 161–175.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2018). Disentangling top-down vs. bottom-up and low-level vs. high-level influences on eye movements over time. *arXiv:1803.07352 [q-bio]*. (arXiv: 1803.07352)

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505–524.

Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, *17*(12), 12:1-35.

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*, *109*(19), 7547–7552.

Sharpe, L. T., Stockman, A., Jagla, W., & Jägle, H. (2005). A luminous efficiency function, V*(λ), for daylight adaptation. *Journal of Vision*, *5*(11), 948-968.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *Information Theory, IEEE Transactions on*, *38*(2), 587–607.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural images statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193-1215.

Smit, A. C., & van Gisbergen, J. A. M. (1990). An analysis of curvature in fast and slow human saccades. *Experimental Brain Research*, *81*(2), 335–345.

Smith, T. J., & Henderson, J. M. (2009). Facilitation of return during scene viewing. *Visual Cognition*, *17*(6-7), 1083–1108.

Snowden, R. J., & Hammett, S. T. (1998). The effects of surround contrast on contrast thresholds, perceived contrast and contrast discrimination. *Vision Research*, *38*(13), 1935–1945.

Solomon, J. A., Felisberti, F. M., & Morgan, M. J. (2004). Crowding and the tilt illusion: Toward a unified account. *Journal of Vision*, *4*(6), 500-508.

Solomon, J. A., & Morgan, M. J. (2000). Facilitation from collinear flanks is cancelled by non-collinear flanks. *Vision Research*, *40*(3), 279–286.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48.

Strang, G., & Nguyen, T. (1996). *Wavelets and filter banks* (2nd ed.). Wellesley, Mass.; Stockport: Wellesley-Cambridge Press.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5), 13:1–82.

Tatler, B. W. (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*(14), 4:1–17.

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), 1857–1862.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, *124*(3), 267–300.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5:1–23.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 1–18.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029–1054.

Teo, P. C., & Heeger, D. J. (1994). Perceptual image distortion. In *IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology* (pp. 127–141). International Society for Optics and Photonics.

Theano Development Team. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, *abs/1605.02688*.

Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, *9*(5), 379–385.

Theeuwes, J., Mathôt, S., & Kingstone, A. (2010). Object-based eye movements: The eyes prefer to stay within the same object. *Attention, Perception, & Psychophysics*, *72*(3), 597–601.

Thibos, L. N., Hong, X., Bradley, A., & Cheng, X. (2002). Statistical variation of aberration structure and image quality in a normal population of healthy eyes. *Journal of the Optical Society of America A*, *19*(12), 2329–2348.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202.

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), 766–786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Trukenbrod, H. A., & Engbert, R. (2014). ICAT: a computational model for the adaptive control of fixation durations. *Psychonomic Bulletin & Review*, *21*(4), 907–934.

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*(1), 507–545.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate bayesian computation.

Journal of Mathematical Psychology, *56*(2), 69–85.

Underwood, G. (2007). Visual attention and the transition from novice to advanced driver. *Ergonomics*, *50*(8), 1235–1249.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*(6), 629–646.

Underwood, G., Foulsham, T., Loon, E. v., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, *18*(3), 321–342.

Ustyuzhaninov, I., Brendel, W., Gatys, L., & Bethge, M. (2017). What does it take to generate natural textures? In *International conference on learning representations.*

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(05), 615–628.

Vecera, S. P., & Farah, M. J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology: General*, *123*(2), 146–160.

Villemonteix, J., Vazquez, E., & Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, *44*(4), 509.

Vincent, B. T., Baddeley, R., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, *17*(6-7), 856–879.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475–494.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

von Helmholtz, H. (1924). *Treatise on physiological optics, translated from the third german edition* (The Optical Society of America, PC Southall, Ithaca, New York, Trans.). Mineola, New York: Dover Publications.

Võ, M. L., Aizenman, A. M., & Wolfe, J. M. (2016). You think you know where you looked? You better look again. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), 1477–1481.

Wade, N. J., & Tatler, B. W. (2011). Origins and applications of eye movement research. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook on eye movements* (pp. 17–43).

Wallis, T. S. A., & Bex, P. J. (2012). Image correlates of crowding in natural scenes. *Journal of Vision*, *12*(7), 6:1-19.

Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, *17*(12), 5:1–29.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Wang, Z., Mohamed, S., & De Freitas, N. (2013). Adaptive hamiltonian and riemann manifold monte carlo samplers. In *International conference on machine learning (icml)* (pp. 1462–1470).

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal*

of Vision, *8*(12), 8:1–13.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Signals, systems and computers, 2004. conference record of the thirty-seventh asilomar conference on* (Vol. 2, pp. 1398–1402). IEEE.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Watson, A. B. (1986). Temporal sensitivity. In Boff & Kaufmann (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 6-1–6-43).

Watson, A. B. (1987). The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, *39*(3), 311–327.

Watson, A. B. (2013). A formula for the mean human optical modulation transfer function as a function of pupil size. *Journal of Vision*, *13*(6), 18:1-11.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740.

Watson, A. B., Borthwick, R., & Taylor, M. (1997). Image quality and entropy masking. In *Electronic Imaging'97* (pp. 2–12). International Society for Optics and Photonics.

Watson, A. B., & Nachmias, J. (1977). Patterns of temporal interaction in the detection of gratings. *Vision Research*, *17*(8), 893–902.

Watson, A. B., & Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, *14*(9), 2379–2391.

Watson, A. B., & Yellott, J. I. (2012). A unified formula for light-adapted pupil size. *Journal of Vision*, *12*(10), 12:1-16.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*. C.F. Koehler.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 681–688).

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., & Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature Communications*, *8*, 14263:1-9.

White, B. J., Kan, J. Y., Levy, R., Itti, L., & Munoz, D. P. (2017). Superior colliculus encodes visual saliency before the primary visual cortex. *Proceedings of the National Academy of Sciences*, *114*(35), 9451–9456.

White, B. J., & Munoz, D. P. (2011). The superior colliculus. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook on eye movements* (pp. 195–213).

Whitney, D., & Levi, D. M. (2011). Visual Crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168.

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination.* (Unpublished doctoral dissertation). University of Oxford.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*(8-9), 1520–1529.

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision*, *10*(4), 6:1–27.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Wilkinson, R. D. (2013). Approximate bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, *12*(2), 129–141.

Williams, D. R., Artal, P., Navarro, R., McMahon, M. J., & Brainard, D. H. (1996). Off-axis optical quality and retinal sampling in the human eye. *Vision Research*, *36*(8), 1103–1114.

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.

Wloka, C., Kotseruba, I., & Tsotsos, J. K. (2017). Saccade sequence prediction: Beyond static saliency maps. *arXiv preprint arXiv:1711.10959*.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). Oxford University Press, USA.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, *15*(3), 419–433.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, *466*(7310), 1102–1104.

Xing, J., & Heeger, D. J. (2000). Center-surround interactions in foveal and peripheral vision. *Vision Research*, *40*(22), 3065–3072.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 201403112.

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 121–134.

Yarbus, A. L. (1967). *Eye movements during perception of complex objects*. Springer.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 3320–3328). Curran Associates, Inc.

Yu, C.-P., Samaras, D., & Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *Journal of Vision*, *14*(7), 4:1-16.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–835.

Zelinsky, G. J. (2012). TAM: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Visual Cognition*, *20*(4-5), 515–545.

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *368*(1628), 20130058.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million

image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Zinkevich, M., Weimer, M., Li, L., & Smola, A. J. (2010). Parallelized stochastic gradient descent. In *Advances in neural information processing systems* (pp. 2595–2603).

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*(1), 41–61.

't Hart, B. M., & Einhäuser, W. (2012). Mind the step: Complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, *223*(2), 233–249.