Sensory Integration under Natural Conditions: a Theoretical, Physiological and Behavioral Approach

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

vorgelegt von

Selim Onat

Osnabrück, Dezember 2009

Acknowledgments

The realization of this thesis wouldn't be possible without the presence of many people who are dear to me. Without Sonja's presence and constant support, especially at moments of weakness, life would have been much more difficult. The time I spent in Osnabrück would have been only half an experience without you. I would like to express my indebtedness to my supervisor Peter König, for his always friendly face, supportive behavior, liberal attitudes, the scientific freedom he gives to his students, and most importantly for creating a working environment that is both friendly and scientifically outstanding at the same time (two qualities rarely simultaneously present). Starting from a single room with a laser printer, building up into a real lab was a most special experience. I would like to also thank my second supervisor Dirk Jancke for the experiments he conducted in Bochum, for his always supporting, friendly attitude and patience. This path was made even more interesting thanks to many friends and colleagues who accompanied me all along. Alper Açık was a precious companion always eager to hold brain-storming sessions and interesting discussions; Cliona Quigley had for me an immense patience for being obliged to listen and correct my horrible written and spoken English mistakes; without Daniel Weiller complex questions would have never been solved so easily, he was always ready to give the good answer; without Johannes Steger it would have not been possible to harbor any positive feelings about Unix based computers and their terminals. I would like to thank also Nora Nortmann, Frank Schumann, Lina Jansen, Wolfgang Einhäuser-Treyer, Saskia Nagel, Boris Bernhardt, Klaus Libertus, Hans-Peter Frey and Sonja Engmann in the NBP group for all their positive presence and professional contributions. A little further away Jörg Hipp, Christoph Kayser, Gudrun Moeller, Rodrigo Salazar, Konrad Körding were helpful colleagues during the first year I spent at INI in ETH Zürich. I would like to thank also distant friends Florian Rigault, Yann Cojan, Manuel Mercier. A big thanks to Jacqueline Griego who has proofread parts of my thesis and Wolfgang Einhäuser-Treyer who accepted to be one of the reviewers of my thesis. Finally I own a big debt of gratitude to my parents who supported me during all my studies. I can not thank them enough for the love they give me.

Curriculum Vitae

SELIM ONAT

Personal Details

Gender: Male
Date of birth: 19th of October, 1978
Place of birth: Istanbul, Turkey
Present Citizenship: Turkish
Languages: Turkish (native), French (near native), English (near native), (German basics)
Address: Albrechtstr. 28, 49069 Osnabrück Germany
Phone: +49 541 969 3509
Fax: +49 541 969 2596
Email: sonat@uos.de
Homepage: www.selimonat.com
Other working experience: Visual Artist, residential VJ in a local night club. Volleyball Player.

Publications[*]

Where do Humans Look at during Scanning of Natural Images: A Comparison of High- vs. Low-Level Features. Selim Onat, Alper Açık, Frank Schumann, Peter König (in Preparation)

***Processing of Locally Presented Natural Movies and Contextual Interactions under Natural Conditions.** Selim Onat, Peter König, Dirk Jancke (in Preparation)

***Processing of Natural Movies Revealed by Voltage-Sensitive Dye Imaging Across Primary Visual Cortex.** Selim Onat, Peter König, Dirk Jancke (2009, Submitted) *PLOS Biology*

***Multiplexing information about position and orientation in early visual cortex: A voltage-sensitive dye imaging study.** Selim Onat, Peter König, Dirk Jancke (2009 Submitted) *PNAS*

***Visual stimulus locking of EEG is modulated by temporal congruency of**

---

[*] Publications which are included in this thesis are marked with an asterisk.

auditory stimuli. Sonja Schall, Cliodhna Quigley, Selim Onat, Peter König (2009) *Exp Brain Res*

Saliency on a natural scene background: Effects of color and luminance contrast add linearly. S. Engmann, T. Sieren, S. Onat, P. König and W. Einhäuser (2009) *Attention, Perception & Psychophysics*

Effects of contrast and its modification on fixation behavior during free viewing of images from different categories. Alper Açık, Selim Onat, Frank Schumann, Peter König (2009) *Vision Research*

How does binocular disparity information influence overt attention. Lina Jansen, Selim Onat, Peter König (2009) *Journal of Vision*

Audio-visual integration during overt visual attention. Cliodhna Quigley, Selim Onat, Sue Harding, Martin Cooke, Peter König (2008) *J of Eye Movement Research*

***Integrating audio-visual information for the control of overt attention.** Selim Onat, Klaus Libertus, Peter König (2007) *Journal of Vision*

Phasic activation of locus coeruleus neurons by the central nucleus of the amygdala. Bouret S, Duvel A, Onat S, Sara SJ (2003) *J Neuroscience*

B.Sc. & M.Sc. Theses Supervised

B.Sc. Theses

Quantitative Analysis of Stable Representations of Disparity in Natural Visual Images by Thesis Sebastian Bitzer

Quantitative Properties of Sparse Binocular Representations of Natural Visual Images by Thesis Markus Goldbach

Cross-modal Integration of Natural Visual and Auditory Stimuli by Thesis Klaus Libertus

The Influence of 2nd Order and Higher Order Correlation in Natural Visual Stimuli on Human Overt Attention by Thesis Sonja Schall

Effects of higher order luminance structure on the selection of fixation points by Thesis Swantje Nadler

Investigation into Differences of Monkey and Human Eye Movements by Thesis Benjamin Auffarth

How Disparity fits into the Statistics and the Neural Processing of Natural Scenes. An EEG Study by Thesis Boris Bernhardt

Human Eye Movements in Three Dimensional Natural Scenes by Thesis Lina Jansen

Baseline Study on Overt Visual Attention by Thesis Anke Walter

A Comparison of Gaze Direction and Subjectively Interesting Points During Free Viewing of Visual Scenes by Thesis Moritz Lehne

Left-Right Asymmetry in Overt Attention by Sarah Mieskes

Integration of Luminance Contrast and Colour Contrast in Directing the Human Gaze by Sonja Engmann

Investigation into Differences of Monkey and Human Eye Movements by Benjamin Auffarth

M.Sc. Theses

Effects of Contrast and its Modifications on Fixation Behavior During Free Viewing of Images from Different Categories by Thesis Alper Açık

Influence of 3D Natural Stimuli on Eye Movements and the Selection of Fixation Points by Thesis Lina Jansen

Integration of Different Features in Guiding Eye-Movements by Thesis Frank Schumann

Audiovisual Integration of Natural Stimuli for the Control of Overt Attention by Cliodhna Quigley

Amplitude Locking of EEG Activity to Dynamic Audiovisual Stimuli by Thesis Sonja Schall

Teaching Experience

Winter Term 2003/2004: Neurobiology of Learning

Summer Term 2003/2004: Neural Coding (based on Dayan and Abbott's Theoretical Neuroscience)

Winter Semester 2004/2005: Numerical Analysis in Neuroscience

Winter Semester 2007/2008: Neural Coding

Education

2004-present
PhD in the Cognitive Science Program at University of Osnabrück, (Germany), under the supervision of Prof. Dr. Peter König

2002-2003
First year of PhD in Institute of Neuroinformatics, ETH/Zürich (Switzerland) under the supervision of Dr. Peter König.

2001-2002
M.Phil. (DEA, Diplome d'études approfondies) in Cognitive Science, Université Pierre & Marie Curie (Paris, France). Thesis (*Étude électrophysiologique de l'interaction fonctionnelle entre le Complexe Amygdalien et le Locus Coeruleus*) accomplished in the Laboratory of Neuromodulation and Memory Processes (Institut de Neuroscience, CNRS) under supervision of Dr. Susan Sara.

2000-2001
M.Sc. (Maîtrise) in Cellular Biology and Physiology (Specialization: Neurosciences), University d'Orsay (Paris, France).

1997-2000
B.Sc. (DEUG + Licence) in Biology (Specialization: Cellular Biology and Physiology), University of Rennes 1 (France).

1996-1997
High School Diploma, French College St. Benoît, Istanbul (Turkey).

References

These persons are familiar with my professional qualifications and my character:

Prof. Dr. Peter König
Thesis supervisor
Email: pkoenig@uos.de
Phone: +49 541 969 2399
Fax: +49 541 969 2596
Address: Osnabrück University
Institute of Cognitive Science
Albrechtstr. 28, 49069 Osnabrueck Germany

Dr. Dirk Jancke
Thesis 2$^{\text{nd}}$ supervisor
Email: jancke@neurobiologie.rub.de
Phone: +49 234 32 27845
Ruhr-University Bochum
Institut für Neuroinformatik
ND 03/70 44780 Bochum, Germany

Dr. (Emerita) Susan Sara
M.Sc. Thesis Supervisor
Email: susan.sara@college-de-france.fr
Phone: +33 1 44 27 14 15
Collège de France
Laboratoire de Physiologie
de la Perception et de l'Action
11, place Marcelin Berthelot
75231 Paris Cedex 05

Abstract

We can affirm to apprehend a system in its totality only when we know how it behaves under its natural operating conditions. However, in the face of the complexity of the world, science can only evolve by simplifications, which paradoxically hide a good deal of the very mechanisms we are interested in. On the other hand, scientific enterprise is very tightly related to the advances in technology and the latter inevitably influences the manner in which the scientific experiments are conducted. Due to this factor, experimental conditions which would have been impossible to bring into laboratory not more than 20 years ago, are today within our reach. This thesis investigates neuronal integrative processes by using a variety of theoretical and experimental techniques wherein the approximation of ecologically relevant conditions within the laboratory is the common denominator. The working hypothesis of this thesis is that neurons and neuronal systems, in the sensory and higher cortices, are specifically adapted, as a result of evolutionary processes, to the sensory signals most likely to be received under ecologically relevant conditions. In order to conduct the present study along this line, we first recorded movies with the help of two microcameras carried by cats exploring a natural environment. This resulted in a database of binocular natural movies that was used in our theoretical and experimental studies.

In a theoretical study, we aimed to understand the principles of binocular disparity encoding in terms of spatio-temporal statistical properties of natural movies in conjunction with simple mathematical expressions governing the activity levels of simulated neurons. In an unsupervised learning scheme, we used the binocular movies as input to a neuronal network and obtained receptive fields that represent these movies optimally with respect to the temporal stability criterion. Many distinctive aspects of the binocular coding in complex cells, such as the phase and position encoding of disparity and the existence of unbalanced ocular contributions, were seen to emerge as the result of this optimization process. Therefore we conclude that the encoding of binocular disparity by complex cells can be understood in terms of an optimization process that regulates activities of neurons receiving ecologically relevant information.

Next we aimed to physiologically characterize the responses of the visual cortex to ecologically relevant stimuli in its full complexity and compare these to the responses evoked by artificial, conventional laboratory stimuli. To achieve this, a state-of-the-art recording method, voltage-sensitive dye imaging was used. This method captures the spatio-temporal activity patterns within the millisecond range across large cortical portions spanning over many pinwheels and orientation columns. It is therefore very well suited to provide a faithful picture of the cortical state in its full complexity. Drifting bar stimuli evoked two major sets of components, one coding for the position and the other for the orientation of the grating. Responses to natural stimuli involved more complex dynamics, which were locked to the motion present in the natural movies. In response to drifting gratings, the cortical state was initially dominated by a strong excitatory wave. This initial spatially widespread hyper-excitatory state had a detrimental effect on feature selectivity. In contrast, natural movies only rarely induced such high activity levels and the onset of inhibition cut short a further increase in activation level. An

increase of 30 % of the movie contrast was estimated to be necessary in order to produce activity levels comparable to gratings. These results show that the operating regime within which the natural movies are processed differs remarkably. Moreover, it remains to be established to what extent the cortical state under artificial conditions represents a valid state to make inferences concerning operationally more relevant input.

The primary visual cortex contains a dense web of neuronal connections linking distant neurons. However the flow of information within this local network is to a large extent unknown under natural stimulation conditions. To functionally characterize these long-range intra-areal interactions, we presented natural movies also locally through either one or two apertures and analyzed the effects of the distant visual stimulation on the local activity levels. The distant patch had a net facilitatory effect on the local activity levels. Furthermore, the degree of the facilitation was dependent on the congruency between the two simultaneously presented movie patches. Taken together, our results indicate that the ecologically relevant stimuli are processed within a distinct operating regime characterized by moderate levels of excitation and/or high levels of inhibition, where facilitatory cooperative interactions form the basis of integrative processes.

To gather better insights into the motion locking phenomenon and test the generalizability of the local cooperative processes toward larger scale interactions, we resorted to the unequalized temporal resolution of EEG and conducted a multimodal study. Inspired from the temporal properties of our natural movies, we designed a dynamic multimodal stimulus that was either congruent or incongruent across visual and auditory modalities. In the visual areas, the dynamic stimulation unfolded neuronal oscillations with frequencies well above the frequency spectrum content of the stimuli and the strength of these oscillations was coupled to the stimuli's motion profile. Furthermore, the coupling was found to be stronger in the case where the auditory and visual streams were congruent. These results show that the motion locking, which was so far observed in cats, is a phenomenon that also exists in humans. Moreover, the presence of long-range multimodal interactions indicates that, in addition to local intra-areal mechanisms ensuring the integration of local information, the central nervous system embodies an architecture that enables also the integration of information on much larger scales spread across different modalities.

Any characterization of integrative phenomena at the neuronal level needs to be supplemented by its effects at the behavioral level. We therefore tested whether we could find any evidence of integration of different sources of information at the behavioral level using natural stimuli. To this end, we presented to human subjects images of natural scenes and evaluated the effect of simultaneously played localized natural sounds on their eye movements. The behavior during multimodal conditions was well approximated by a linear combination of the behavior under unimodal conditions. This is a strong indication that both streams of information are integrated in a joint multimodal saliency map before the final motor command is produced.

The results presented here validate the possibility and the utility of using natural stimuli in experimental settings. It is clear that the ecological relevance of the experimental conditions are crucial in order to elucidate complex neuronal mechanisms resulting from evolutionary processes. In the future, having better insights on the nervous system can only be possible when the complexity of our experiments will match to the complexity of the mechanisms we are interested in.

Contents

1 General Introduction 1

1.1 Context 1

1.1.1 Representationism in Cognitive Science 3

1.1.2 About this Thesis 6

2 A Simulation Approach: Learning Binocular Receptive Fields from Natural Movies 11

2.1 Context 11

2.2 Introduction 12

2.3 Methods 14

2.3.1 Recording of Videos 14

2.3.2 Building the Stimulus Set 14

2.3.3 Structure of the Neuronal Network 15

2.3.4 The Goal Function 16

2.3.5 Analysis of Receptive Fields 17

2.4 Results 18

2.4.1 Orientation and Spatial Frequency Selectivity 18

2.4.2 Disparity Encoding 19

2.4.3 Translation Invariance 21

2.4.4 Binocular Dominance 21

2.5 Discussion 21

2.5.1 Why Do Receptive Fields Have the Structure They Have? 21

2.5.2 How Could Stable Features Be Extracted by Neurons? 23

2.5.3 Relation to Sparseness 23

3 An Optical Imaging Study: Cortical Dynamics During Processing of Natural Movies 25

3.1 Context 25

3.2 Analysis of Optical Imaging Data 26

3.3 Processing of Natural Stimuli in V1 27
  3.3.1 Introduction 27
  3.3.2 Experimental Procedures 28
    3.3.2.1 Stimulus Acquisition and Presentation 28
    3.3.2.2 Stimulus Conditions 29
    3.3.2.3 Experimental Setup 29
    3.3.2.4 Retinotopical Mappings 29
    3.3.2.5 Data Acquisition 30
    3.3.2.6 Preprocessing 30
    3.3.2.7 Region of Interest Selection 31
    3.3.2.8 Latency Detection 31
    3.3.2.9 Singular Value Decomposition 31
    3.3.2.10 Detection of Confidence Intervals 31
  3.3.3 Results 32
    3.3.3.1 Spatio-Temporal Population Dynamics Evoked by Natural and Grating Stimuli 32
    3.3.3.2 Locking of Activity to the Stimulus Motion 34
    3.3.3.3 Spatio-Temporal Separability of Activity Dynamics . . . 35
    3.3.3.4 Singular Value Decomposition Unravels Two Overlapping Maps. 37
    3.3.3.5 Comparison of Evoked Activity Levels 39
    3.3.3.6 Specific vs. non-Specific Activity 41
    3.3.3.7 Differences in the Time-Courses of Activities 43
    3.3.3.8 Effect of Variability of Local Luminance Contrast 44
    3.3.3.9 Deceleration/Acceleration Notch 44
    3.3.3.10 Strength of the Acceleration vs. Orientation Selectivity . 46
    3.3.3.11 Population Sparseness 46
  3.3.4 Discussion 47
3.4 Contextual Effects under Natural Conditions 53
  3.4.1 Introduction 53
  3.4.2 53
  3.4.3 Results 54
    3.4.3.1 Locally Presented Natural Movies Evoke Localized Dynamic Activity Patterns 54
    3.4.3.2 Long-range Cortical Connections Reduce Response Time 55

3.4.3.3 Lateral Connections Exert Facilitatory Influences under Natural Conditions 55
3.4.3.4 Facilitatory Effects are Sensitive to the Spatio-Temporal Coherence of the Stimulus 58
3.4.4 Discussion 60

4 An EEG Study: Motion Locking and Crossmodal Integration in Humans 62

4.1 Context 62
4.2 Introduction 63
4.3 Experimental Procedures 65
4.3.1 Participants 65
4.3.2 Experimental Design 65
4.3.3 Stimuli 66
4.3.4 Presentation & Recording 67
4.3.5 Behavioural Analysis 67
4.3.6 Pre-processing & Artefact Removal 68
4.3.7 Stimulus Locking 68
4.3.8 Evaluating Statistical Significance 70
4.4 Results 70
4.4.1 Task Performance 71
4.4.2 Stimulus Locking: Can We Find it in Human EEG? 71
4.4.3 Visual Locking 71
4.4.4 Auditory Locking 74
4.4.5 Is there Evidence for Multisensory Interactions? 74
4.4.6 Do the Effects Change over Time? 76
4.5 Discussion 77

5 A Behavioral Approach: Crossmodal Integration for the Control of Overt Attention 82

5.1 Context 82
5.2 Introduction 83
5.3 Materials & Methods 84
5.3.1 Participants and recording 84
5.3.2 Stimuli 85
5.3.3 Experimental paradigm 85
5.3.4 Data analysis 87

5.3.5 Probability distributions 87
5.3.6 PDF parameterization 87
5.3.7 Modeling the cross-modal interaction 89
5.3.8 Luminance contrast measurements 90
5.4 Results 90
5.4.1 Subjects' Gaze Is Biased towards the Sound Location. 90
5.4.2 Effect is More Prominent in the First Half of Presentation 91
5.4.3 Specific Interaction of Visual and Auditory Information 92
5.4.4 The Integration Is Linear 96
5.5 Discussion 100

6 General Discussion 104

7 Appendix 108

7.1 Description of Optical Imaging Data & Generic Processing Steps: From Photon Counts to Neuronal Activity 108
7.1.1 Some Terminology 108
7.1.2 Properties of Raw Data 109
7.1.3 Frame-zero Correction by Division 110
7.1.4 Why Division? 111
7.1.5 Blank Correction (Fractional Change) 111
7.2 Behavioural Pilot Study 115

Disclaimer 116

Abbreviations 116

Bibliography 117

List of Figures

1.1 Embodied Cognitive Science 7

2.1 Recording of Natural Movies. 15
2.2 The Structure of the Artificial Neuronal Network. 16
2.3 A Selection of RFs Representing Optimally Stable Representations of Stereoscopic Natural Movies. 19
2.4 Feature Selectivity of the Simulated Complex Cell Population. 20

3.1 Retinotopical Mapping. 30
3.2 Recording Cortical Responses to Natural Stimuli and Gratings. 33
3.3 Motion Locking of Average Activity. 35
3.4 Separability of Spatio-Temporal Activity Dynamics. 36
3.5 Decomposition of Evoked Cortical Voltage-Sensitive Dye Responses to Moving Gratings. 38
3.6 Propagation of Activity across Stationary Orientation Maps. 40
3.7 Propagation Speed of Cortical Activity. 41
3.8 Differences in the First-Order Characteristics of Response Levels. 42
3.9 Transient Part of the Response to Stimulus Onset. 45
3.10 Second and Higher Order Characteristics of Response Distributions. . . . 47
3.11 Presentation of Natural Movies. 54
3.12 Impact of Coherent Context on Cortical Activation Patterns. 56
3.13 Effects of Contextual Stimulation on the Latency of Activation. 57
3.14 Distant Natural Stimulus Exerts Facilitatory Effects on Local Activity Levels. 58
3.15 Magnitude of Facilitation Depends on Stimulus Coherence. 59

4.1 Schematic Diagram of Both Experimental Paradigms. 66
4.2 Two Correlation Analysis Approaches. 69
4.3 Stimulus Locking of EEG Power to Visual Stimuli. 72
4.4 EEG Waveforms Lock to Visual Input at Multiple Sites. 73

4.5 No Stimulus Locking to Auditory Input. 74
4.6 Temporal Progression of Stimulus Locking. 75

5.1 Experimental Paradigm and Subject Selection. 86
5.2 Bias of Subjects' Gaze towards the Sound Location. 91
5.3 Distribution of Fixation Density Shifts towards Sound Location. 92
5.4 Effect of Localized Auditory Stimulus is More Prominent in the First Half of Presentation. 92
5.5 Specific Interaction of Visual and Auditory Information. 93
5.6 Auditory and Visual Information are Integrated. 95
5.7 Prominent Contribution of Unimodal Saliencies. 96
5.8 The Integration is Linear. 98

7.1 Raw Data Directly Output from the Camera. 109
7.2 Fluctuations in the Basic Statistics Extracted from the Raw Data of a Given Experiment. 110
7.3 Relationship between Prestimulus Average Luminance and Modulation Depth of a Pixel. 112
7.4 Effect of Frame-zero Correction. 112
7.5 Effect of Frame-zero and Fractional Change Operations on the Fluctuations.113

1

General Introduction

1.1 Context

One of the most surprising facts about life is its diversity. Besides the biological diversity of species which is in the focus of classic and modern biology across centuries, another aspect that is much less investigated is just as worth being surprised: *The diversity of the experienced world.* The way living organisms experience the world is also subject to extreme diversity. Naturally here, there is no possibility for any claim of this order to be supported by a rigorous scientific proof because on the one hand, we have no practical access to how other animals may be experiencing the external world and on the other hand, we simply have no idea of how and why we, as humans, do experience the world as we do (O'Regan & Noë 2001). However, given the enormous diversity of animals and their sensory organs, one may make the claim that the world as it would potentially be experienced by living beings is infinitely complex (Freeman 2000). As a matter of fact we, animals, do not experience this infinity but only a part of it. And interestingly different animals do it differently. What is the origin of this diversity? More precisely, what determines that a sensory system should be sensitive to a given set of aspects of the environment and ignore, to a good extent, others?

One answer, which is certainly the Modern Age's most prevailing one, would be that these diverse selectivities result from the clockworks of evolution shaping different sensory organs and epithelia in order to *represent* different and specific aspects of external world by means of genetic and/or epigenetic mechanisms. Each unique sensory domain could then be seen as different pair of goggles directed to different parts of the same reality, which is the environment. Said another way, this view has a specific answer to the popular chicken-egg problem: Which one, the sensory system responsible for sensation or the environment, did come first? According to this view, the environment should have been there before the sensory system evolved to process and represent the incoming information by virtue of natural selection (Varela et al. 1992). Therefore this account of sensory diversity presupposes that the world has *pregiven* properties which are accessible to the organisms directly. It assigns a totally passive observer status to the organism with the absolute capacity of having access to different properties of the environment.

But here we should stop and ask what an incoming signal is . Does the environment let itself be experienced by observers without any constraints? The answer is most certainly negative given the fact that the input to the nervous system depends not only on the environment but also on the animal's more general properties such as its body shape, locomotion, speed, head shape, inter-ocular distances, dexterity among many more traits. The obvious relationship of the body shape and multitude of potentials it generates or restricts, suggests that the problem of sensory diversity is not a simple problem which could be explained by a representation of the external world and further attunement of these representations. One should recognize the primacy of animals being animated beings. As a fact of definition, animals are self-animated beings and the capacity of self motion is their unshared trait among all living beings. Importantly, the nervous system is an invention of the Animal Kingdom: that is to say that the link between action and sensation is created via the nervous tissue.

Strikingly, a representationist account of sensory diversity does not make recourse to this defining property of animal nature. Hence it ignores to a large extent their dynamic nature. It simply focuses on the sensation and creates a simple picture of the relationship between the environment and the organism. Such a view reduces the animal to a disembodied agent where it simply receives signals from the external world, to which he has full access from the beginning. However considering the organism in its embodied context, we accept the fact that what reaches central sensory neurons primarily depends on, the environment, and also importantly, how motor neurons make use of the very specific body, which in turn depends on the very shape of the body and the manner in which it interacts with the environment. Therefore it becomes increasingly harder to defend the position of a pregiven world that is represented by the organism without taking into account the shape and design of the animal. The shape and design of the animal specifies the sensory motor space of the animal; these are the aspects which an *embodied* science of mind can not disregard (Clark 1999). Let's illustrate this view with an excellent passage of *The Embodied Mind* (Varela et al. 1992):

> It is well known that honey bees are trichromats whose spectral sensitivity is shifted toward the ultraviolet. It is also well known that flowers have contrasting reflectance patterns in ultraviolet light. Consider now our "chicken-and-egg" question [...]: Which came first, the world (ultraviolet reflectance) or the image (ultraviolet sensitive vision). Most of us would probably answer with little hesitation, The world (ultraviolet reflectance). It is therefore interesting to observe that the colors of flowers appear to have *coevolved* with the ultraviolet sensitive, trichromatic vision of bees.
>
> Why should such coevolution occur? One the one hand, flowers attract pollinators by their food content and so must be both conspicuous and yet different from flowers of the other species. On the other hand, bees gather food from flowers and so need to recognize flowers from distance. These two broad and reciprocal constraints appear to have shaped a history of coupling in which plant features and the sensorimotor capacities of bees coevolved. It is this coupling, then, that is responsible for both the ultraviolet vision

> of bees and the ultraviolet reflectance patterns of flowers. Such coevolution therefore provides an excellent example of how environmental regularities are not pregiven but are rather *enacted* or brought forth by a history of coupling."

This example emphasizes an important concept of the so-called *enactivist* point of view: *structural coupling*. In this example, bees and flowers are two components of a system. These, by virtue of their coupling codetermine the fate of each other. In this paradigm, the idea of a world with pregiven attributes that the agent needed to represent does not occupy a central position. Admittedly the agent needs a sensorial system that reacts to the external world, but the principle characteristics of the sensory-motor system are not organized for the representation of the external world. Both the environment and the agent are embodied within a dynamic system that enacts the observed visual sensitivities of bees. Bees do not need to know the exact properties of the external world. They don't need to create a sensory system which faithfully represents the world. It suffices for the subject to distinguish the relevant objects from the background of meaningless stimulation. This example shows that the *representationism* (at least at its strong definition) is not a necessary component of the biological and sensorial diversity generation. From this perspective, the fact that different animals have different sensitivities with respect to the environment results from their specific history of structural coupling, which endows them with basic capacities for solving problems related to their survival.

1.1.1 Representationism in Cognitive Science

Interestingly, representationism (takes the label of *Cognitivism* in the field of Cognitive Science) translates one-to-one to the field of Cognitive Science and constitutes the major paradigm and working hypothesis. Here also, representationism is based on the very assumption that the world has pregiven properties which are then represented by the mind in order to solve different tasks or to produce intelligent behavior. Accordingly, any intelligent behaviour finds its sources on the computations which are done on the symbolic level based on the representations of an external world. Clearly today, representationist accounts of the intelligent behaviour can be found in many different contexts. It wouldn't be erroneous to assert that this view occupies the defining zeitgeist in Cognitive Science.

Let's illustrate this paradigmatic view with a real-world example and take the task that an outfield baseball player has to solve. The question of interest here is to understand how outfield players in baseball do know where to run, starting from the point where the ball hits the bat. Given the high speed levels the ball reaches and the large distances between the ball and the outfield player, it is reasonable to assume that the visual cues, such as for example stereoscopic depth, are not reliable. Therefore the question of how this is achieved raises interest.

Classically this task was thought to be solved by humans by realizing unconscious complex mathematical computations in order to derive the trajectory of the ball based on variables extracted from the external world such as for example its speed, acceleration (McLeod & Dlenes 1993). McBeath et al. (1995) provided a different account that relies on the idea that when outfield players start running to catch the ball, they choose a trajectory for their own body which linearizes the trajectory of the ball according to the optical information they receive. This is an important switch in our conceptualization of the intelligent behaviour as it removes any need to extract parameters from the external world, and thus, the realization of complex mathematical computations on them. The problem raised by fact that we are not consciously aware of these computations (see Mind-Mind problem (Jackendoff 1987)) is also eliminated. Rather it relies on sensory-motor coupling and puts the body into a central position by involving the active motor behaviour of the subject: the player using his sensory-motor capacity coupled dynamically to the external world can exhibit the required intelligent behaviour without actually representing any external parameters.

It is generally within the representationist paradigm that Computer Scientists "inspired" by biological vision are interested in reconstructing the three-dimensional world based on two-dimensional, and thus ambiguous, retinal images. For example, a chair may have extremely different two-dimensional projections onto a two dimensional surface depending on the angle of view. Here it is of extreme interest to create a visual algorithm which is able to understand the content of images. Although this task is extremely easy for the primate's nervous system, computer algorithms are rather poor in their recognition performance.

A typical Computer Vision approach for solving this task would be to design an algorithm for recognizing objects based on their two-dimensional images. A dataset consisting of images taken from a camera, which passively creates images, would be the major source of information in order to solve this task. A classifier trained on these data would learn how to categorize different images depending on the presence or absence of the target object. Clearly as shown in this example, the representationist approach ignores the situatedness of biological intelligence, its dependence on a body and also on a history of sensory-motor coupling which underpins the performance of living system, therefore it reduces the visual system to a passively observing camera. As criticized in *Action in Perception* (Noë 2004), there is simply no fundamental reason to think that the data which carries the information about the external world is contained exclusively in the static two-dimensional retinal images. However, the outlined approach in computer science is completely static, passively observing the outside world and radically disembodied. It wouldn't be erroneous to assert that since the approach outlined by David Marr in *Vision* (Marr 1982), the paradigm in Computer Vision, one of the most prominent fields in Cognitive Science, has not made significant progress toward incorporating embodied aspects of biological vision.

In experimental fields of Cognitive Science however the importance of the active bodily participation was very early recognized (even though here also it never became the major paradigm). One of the classic behavioral experiments which underlines the primacy of an active bodily involvement for the development of perceptual skills is the work of Held & Hein (1963). They showed that a well functioning central nervous system requires the active motor involvement of the animal during its developmental stages. In their now classical experiments, they deprived a group of cats (cats are mammals commonly used in physiological experiments investigating sensory systems) of the active motor component of their behavior while they were exposed to the same visual input patterns as another group of cats which were not constrained in their active behavior (see Fig. 1.1A for a schematic view of the experimental setup). According to the representationist view, as both groups are being exposed to the same visual input, the development of their visual system should be at least comparable in their performances following the restricting rearing period. However the restrained cats were severely impaired in their basic perceptual and motor skills. Their results showed that passive observation of the outside world was not sufficient for generating basic sensory-motor capacities used to solve simple tasks.

Much more recently the FeelSpace® (Nagel et al. 2005) project conducted in the Neurobiopsychology Department in Osnabrück is based on a very similar idea. Within this project, humans learn how to make use of a mechanical belt by carrying it during their everyday life (Fig. 1.1B). By means of a compass, the belt vibrates always at the North direction independent of body direction. This results in the vibration of different parts of the body in contact with the belt, depending on the orientation of the subject in the North-South axis. Initially the stimulation by the belt is rather simple and meaningless, however the tactile stimulation gains a biological meaning within a relatively short temporal interval (in the order of weeks) through the building up of a sensory-motor coupling history resulting from the interaction between the subject and the effect of the externally plugged device. Not only that, subjects acquire the capacity of incorporating this ”new sense” on navigational tasks they are required to solve, additionally their phenomenal perception of the space becomes different, if not extended. For example, they report awareness of the orientation of their dining table with respect to the position of the orientation of their office desk. This fact lies indeed on a major divergence point between enactivist and representationist approaches. Whereas the representationist perspective has serious difficulties with the incorporation of new senses to an old body, the enactivist view predicts that the emergence of new senses should not be given a special status considering that this is how we make use of our biological sense apparatus during development in any case.

Given the importance and necessity of active bodily participation for the normal development of a visual system, one might still ask how humans solve visual tasks while they are not bodily active: for example, in simple laboratory conditions. Suppose we are given the task of predicting the success (whether it's “in” or “not”) of a basketball player's shot and provided necessary video material recorded under controlled conditions while players are aiming the ball to the basket (Fig. 1.1C). How do humans solve this task. Given that we are presented pixel-based video data, one approach would be to follow a strategy of finding some heuristics based on the visual data. For example initial ball trajectory, fore-arm angle etc. would be the best candidate parameters with predictive value regarding the ball's trajectory. If this were true, this task could exclusively be solved by involvement of visual areas in the central nervous system.

Using transcranial-magnetic stimulation and measuring the muscular evoked potentials, Aglioti et al. (2008) tackled this issue experimentally. What they found was evidence for the large-scale involvement of the central nervous system, including motor cortex, during this task. The activity of neurons in the motor cortex corresponding to the arm area (the only area tested) were found to be involved during this task, suggesting that subjects were using the parts of their brain usually involved during such tasks in real-life. Importantly, people who were not experienced were not able to solve this task, suggesting that what is to be perceived is not a process which relies on symbolic inference but rather depends on previous bodily experience. This interpretation directly supports an embodied action view of the nervous system, which is valid even during conditions that do not require the usage of a body.

The dominant view and scientific enterprise in the Cognitive Science was thus far dominated by representationism. To this view is an intrinsic difficulty of conceiving the organisms as being embodied agents associated. Recent advances in the field of Cognitive Science make it unlikely that we can achieve an understanding of the intelligent behavior without conceiving the organism's dynamic nature and its situatedness in the environment. In the future this will necessarily result in a shift of our current computer-mind metaphor toward a more dynamic environment-body metaphor.

One point must be considered carefully in the discussion of representationism and taken care to note that the point here is not to reject the existence of the representations in the brain. Most certainly the activity of neurons could match to some external variables faithfully. And as observers we may assign to these neurons a representative role. Neurons within the motor cortex could very well represent different movement directions of the arm (Georgopoulos 1994); the activity of neurons responsive to visual stimulation may very well correspond to, under certain restricted conditions, different parameters of the visual stimulation. For example, the activity of middle temporal area neurons are known to match to the speed of the stimulus (Liu & Newsome 2006) and the stimulation of these neurons can cause perceptual shifts on the monkey's perception in a predictable way in accordance with the feature selectivity of these neurons (Celebrini & Newsome 1994). But we should be aware that given this knowledge we can not deduce that the nervous system as a whole is operating on the basis of representations. The main point is to emphasize the fact that sole observation of a correlation between neuronal activity and parameters of external world do not allow us to conclude that the cognition and intelligent behaviour is based on manipulation of symbols extracted from outside world. It is the representationism which is problematic and not representations.

1.1.2 About this Thesis

How would it be possible to study the physiology of sensory systems within an embodied paradigm? One strategy would be to study sensory systems under conditions as close to the real-world operational conditions of the animals as possible. This requires that we, as experimentalists, need to be aware of the natural embodiment of the animals, and try to develop an understanding of the role of sensory neurons within a context that makes sense for their situatedness.

For this reason, during physiological investigations where I focused on the visual system of the cat, I aimed to approximate as closely as possible the visual signals that cats receive during their real-world conditions (Fig. 1.1D). To this end, I recorded stereoscopic natural movies from their perspective using a pair of micro-cameras carried on the head of freely moving cats viewing their natural habitat. A large part of the theoretical and experimental results presented in this thesis relies heavily on these movies. In order to avoid technical complications due to eye movements, natural movies were recorded in head centered coordinates. The fact that cats do not make large saccadic movements during active behaving (Einhäuser et al. 2008) makes it likely that the movies recorded by cameras do represent the input to the sensory system within reasonable limits.

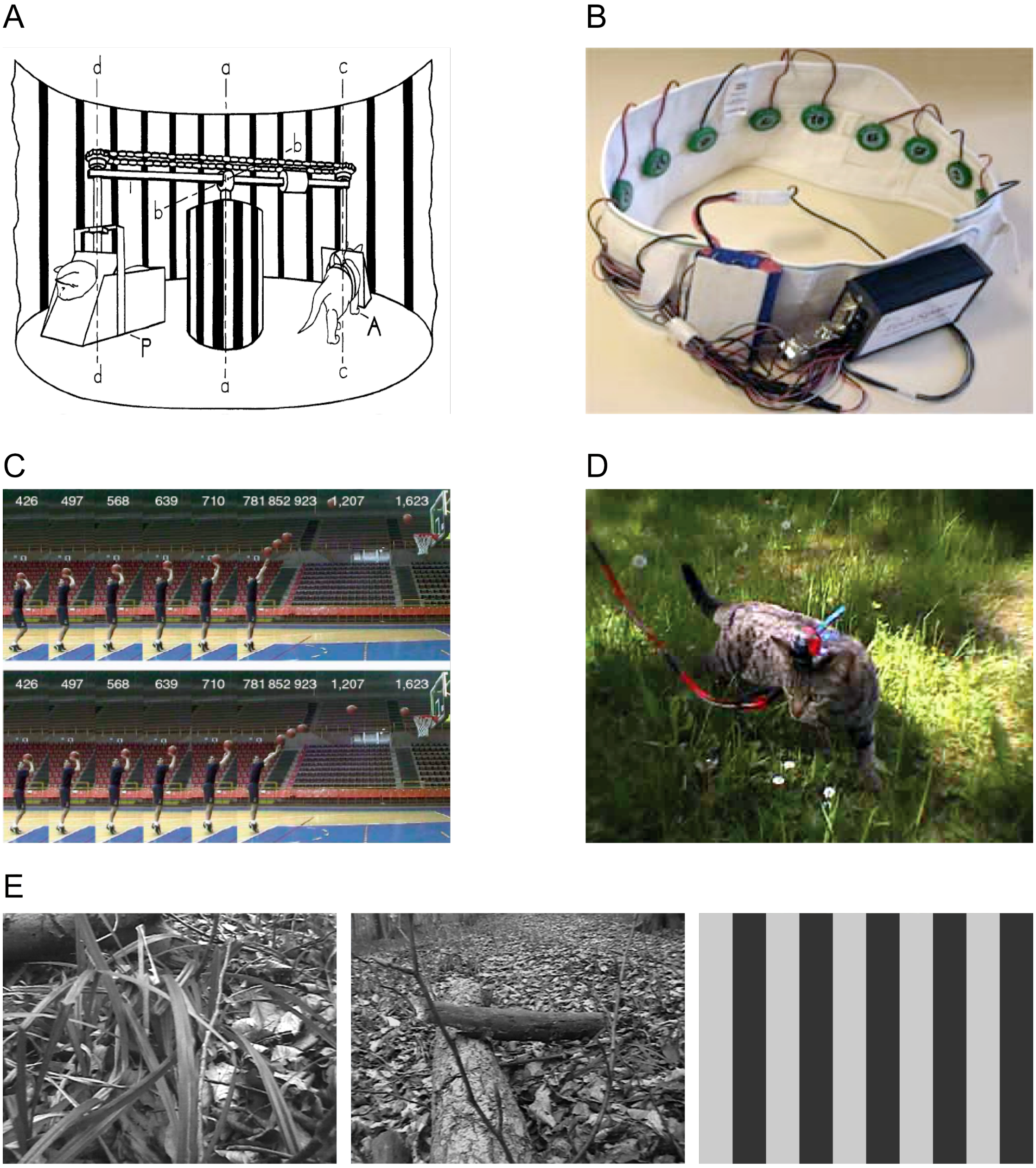

Figure 1.1: **Embodied Cognitive Science. (A)** Apparatus for equating motion and consequent visual feedback for an actively moving (A) and a passively moved (P) S[ubject] (Fig. and caption from Held & Hein (1963), my text is indicated within square brackets. **(B)** One of the earliest models of the FeelSpace® belt. (Image taken from www.feelspace.de). 13 vibrators provide tactile information depending on the positioning of the subject along the North-South axis. **(C)** Fig. taken from Aglioti et al. (2008). Showing the visual information that human subjects were provided during a task which consisted of predicting whether a sequence of frames representing a basketball player shooting would be a success or fail. The frames were cut shortly before or after the ball left the hand of the player. **(D)** Making of stereoscopic natural movies by freely moving cats in a natural habitat. A pair of cameras carried by a cat was connected to two VCR recorders carried by the experimenter. (See also Fig. 2.1 for more detail.) **(E)** An example of two frames extracted from movies are shown together with an artificial stimulus. These kind of stimuli were used in our physiological and modelling experiments.

Since the discovery of simple and complex cells in the primary visual cortex (V1*) by Hubel and Wiesel (Hubel & Wiesel 1959; 1962a), which brought them the Nobel Prize in 1981 for the work they carried out during early 60's, the scientific investigations of sensory systems are dominated mainly by the usage of simplified stimulation protocols in physiological experiments. In visual experiments usage of drifting gratings, random dots, flashed bars and moving edges were (and are still today) very common for the characterization of the receptive fields (RF). There is no doubt that the spatial and temporal properties of such visual images and movies are extremely unusual and disparate with respect to images encountered in the natural environments (see for example Fig. 1.1E) from the animals perspective. While a natural stimulus such as, a photograph taken in a wood, contains many different spatial frequencies and orientations simultaneously, simplified stimuli are generally constituted of a single orientation and spatial frequency at a time. The cortical responses to natural and complex stimuli were marginally investigated by only a few experimenters (Creutzfeldt & Nothdurft 1978).

It is no doubt that technological restrictions during these early times of electrophysiology favoured the usage of simplified stimuli during experiments. However, beside these technological constraints, having an easily controllable and parameterizable source of stimulation was also seen as a major benefit. Therefore one can say that conceptual drives had also a significant contribution in increasing the popularity of simplified stimuli. Nevertheless, during this initial period of the sensory neuroscience, usage of such simple stimuli ensured a constant accumulation of data, results and ideas. Different theories (for example related to the emergence of orientation selectivity) which are debated even today, were proposed during these initial times. Therefore, the very basics of our current knowledge on sensory processing relies largely on these experiments realized under reduced stimulation settings. Consequently little is known about how sensory neurons process complex natural inputs. Furthermore we currently are only partially starting to understand whether and how cortical neurons are evolutionary adapted to process their most common input patterns. Therefore natural stimuli form the best choice of stimuli if the search is for the existence of more complex and sophisticated cortical mechanisms related to the processing of real-world input.

However, this view does not constitute a consensus among specialists in the field. Rust & Movshon (2005) argue that the physiological systems can be studied by simplified, parameterized stimuli forming the constitutive parts of more complex signals such as natural images. This view however has one implicit assumption: namely, it asserts that the knowledge gathered regarding the functioning of neurons with experiments where simple stimuli is used, can be transferred without any complication to the cases involving more complex situations such as processing of real-world inputs. As of today, we do not have enough data to clearly answer whether this position is wrong or not. I think however that the results presented in this thesis (Chapter 3) contributes to this topic.

As noted at the very beginning, the sensory input that animals are exposed to are constrained in many respects. First of all the natural habitat enormously restricts the possible set of input patterns, that is to say not all signals are equally likely (Simoncelli & Olshausen 2001, Barlow 1961a, Atick et al. 1992). Consider for example a set of 10 by 10 pixel-image representing the light falling within a restricted area of the retina with 255

---

*complete list of abbreviations are given at the end of this thesis.

discrete numbers of grays. It is trivial to show that the total number of all possible inputs is simply astronomical. However, studying statistical properties of natural images, one realizes that real-world input spans only a very limited subspace (Chandler & Field 2007). It is therefore reasonable to consider the possibility that sensory neurons developed adaptational skills in order to react optimally to such stimuli. Such an adaptational strategy has many advantages. First of all, restricting the processing power to a very limited subspace may increase reaction times in face of life-threatening situations where the survival is important. Moreover, recently it has been proposed that the energy consumption may be an important constraint for the development and functioning of the nervous system (Laughlin et al. 1998). From this perspective RF which minimize the total number of spikes by coupling to the environment as efficiently as possible are of biological interest as they reduce the required energy consumption for the transmission of information.

The fact that natural images spans only a limited portion of all possible stimuli means that the real-world input that the animal receives contains certain regularities. A milestone study by Laughlin (1981) showed that the contrast sensitivity of neurons in large monopolar cells in the fly visual system follows the distribution of contrast values encountered in the habitat of the fly. The incremental change in spiking activity following an incremental change in the contrast of the stimulus was not random nor constant. The change in number of spikes was minimal for contrast values which were rare in the environment and maximal for the contrast levels which were most common. This specific contrast sensitivity curve has the consequence that the neurons in the visual system of the fly are most sensitive to resolving the most common signals. Additionally from an information theoretical point of view, this leads to a distribution of spiking activity having the maximum rate of information because all possible spike counts occur with equal probability in the long run. It is thought that this specific way of translating luminance contrast values into spiking activity is the result of an adaptive process leading to the efficient processing of natural input. In the visual system of mammals such adaptive phenomena has also been observed. In a now classical study, Dan et al. (1996) recorded responses of neurons located in the lateral geniculate nucleus (LGN) of cats. This nucleus is the target of the retinal ganglion cells and is located only one synapse before V1; it therefore stands in the mid-way between retina and cortex. They showed that the spatio-temporal RF properties of these neurons are specifically designed to process natural input efficiently so as to maximize the information content of the spike trains. Using another type of stimuli with properties deviating significantly from the statistics of natural images, such as white-noise patterns, they observed that the responses of neurons were much less equally distributed: meaning that they were inefficiently conveying the information. They interpreted their results as being a consequence of an adaptational strategy of neurons to natural real-world input.

The statistical properties of the input signals that reach neurons in the early cortical areas are not constrained solely by external world. They also depend on the animal's body shape, head movements, eye movements and on the coordination of the same. Therefore in addition to spatial properties of natural images, temporal properties also may in principle be equally important. That is to say that the input received from the beginning on by the sensory neurons about the external world is dynamic and very idiosyncratic with respect to the animal in question. In the first part of this thesis (Chapter 2), I aim to understand how spatial properties of binocular RFs located in area 17/18 (A17/18) of cats and V1 of monkeys are related to the spatio-temporal statistics of natural signals received by the visual system. The spatio-temporal properties of the natural movies I used here are constrained by, both the spatial properties of natural images and the bodily motion of cats recording these movies. I employed an unsupervised learning scheme in an artificial neuronal network in order to derive short-hand mathematical descriptions of learning principles that lead to the observed binocular complex and simple cells RFs. The learnt RFs do possess striking similarities to the neuronal RFs observed in physiological experiments.

In the second part (Chapter 3), I use a state-of-the-art neuronal recording method, namely voltage-sensitive dye imaging (VSDI), in order to unravel the dynamics of large numbers of neuronal populations during exposure to naturalistic movies and I compare these dynamics to the activity levels evoked by simplified laboratory stimuli (Section 3.3). This is the first time that a large-scale recording method directly measuring neuronal activity has been used in conjunction with natural stimuli to investigate neuronal dynamics at the mesoscopic level. My results demonstrate that the cortical processing structures are indeed adapted to their real-world input and the artificial stimulation paradigms, using for example, moving bar stimuli, lead the system towards an operating regime which is different than natural movies. Moreover, in the last part of the same Chapter (Section 3.4), by presenting natural movies locally through either one or two apertures, I analyzed how the presence of contextual information influences the local processing. My results show that the dense intra-cortical connectivity patterns are at work under conditions of complex dynamics stimulation.

In the third part of this thesis (Chapter 4), inspired by the dynamic nature of the natural movies, I investigate the processing of dynamic stimuli and the effect of dynamic extramodal sensory signals in humans. Previous results in our working group (Kayser & König 2004) showed that dynamic stimuli such as movies recorded by cats induces modulations in the local field potentials (LFP) at different frequency bands, a phenomenon called motion locking. Due to the fact that the VSDI is not well-suited for recording of high frequency oscillations, we recoursed to the EEG which is a well-known and established recording method. I investigated how such dynamic stimuli is processed in humans and the effect of contextual auditory information that was either congruent or incongruent with the visual input. Our results showed, for the first time in humans, the existence of motion locking using the method of EEG thus extending the results obtained in the cat cortex. Moreover I provided evidence that motion locking in the early visual areas is subject to modification in the presence of contextual auditory signal if it is congruent to the motion of the visual stimulus.

In the fourth and last part of this thesis (Chapter 5), the active behavior of humans during real-world input was in my focus. In order to complement the physiological findings reported above concerning the existence of short and local integrative mecanisms at the cortical level, I measured eye movements in human subjects as they were freely observing naturalistic photographs and investigated how a localized auditory signal influenced their behavior under natural conditions. My results showed that the presence of localized sound stimuli shifted the eye movements toward the side of the auditory stimulus; however, this was more than a simple orientation behavior and reflected the outcome of an integrative phenomenoun taking place during the planning of eye movements.

2

A Simulation Approach: Learning Binocular Receptive Fields from Natural Movies*

2.1 Context

One of the fundamental questions in current theoretical neuroscience is how to understand the rules governing RF properties. Neurons gather their functional properties by virtue of the specific connectivity pattern on the surface of their dendrites with the incoming axons. A neuronal subtype called *simple cell*, found mainly in the thalamorecipient layers of early visual cortex, do have RFs with elongated ON and OFF subfields endowing them with the ability to detect edge-like structures (Hubel & Wiesel 1959). Compared to the circular RFs of retinal ganglion (Kuffler 1953, Barlow 1953) and LGN (Hubel & Wiesel 1962a) neurons, this is considered as an operation resulting in sensitivity to more elaborated features. The other major subtype, *complex cells*, found mainly in the superficial layers (Martinez et al. 2005), are in addition to be sensitive to the orientation of the edge, they possess the intriguing property of being relatively invariant to its precise position within the RFs. Neuronal computations leading to invariance with respect to a given visual feature such as for example the position of the stimulus, are ubiquitous in the ventral visual system (Ito et al. 1995). For example neurons in the fusiform gyrus (Kanwisher et al. 1997) are, beside being selective for faces, relatively invariant to the angle of view. From this perspective the complex cells we are investigating here occupy a well-defined study case for the understanding of learning and emergence of invariances. What are the computations realized in the dendrites of these neurons, and most importantly, can we derive compact mathematical descriptions of the learning principles for these operations?

The spatial properties of complex cell RFs are compatible with the hypothesis that

---

*I am thankful to Christoph Kayser for his support for the preparation of the results presented in this Chapter.

cortical neurons form *optimally stable* representations of time-varying real-world input (Hyvärinen & Hoyer 2001, Einhäuser et al. 2002, Körding et al. 2004, Wiskott & Sejnowski 2002). Within this scheme an artificial neuronal network is fed with natural input and the response of the neurons are optimized with respect to an objectively defined constraint. This optimization forces neuronal activities to change as slow as possible across time in response to a representative set of natural image sequences by virtue of the specific organization of the learnt connectivity weights. It has been shown that extracting optimally stable features from natural movie sequences in conjunction with a basic neuronal model leads to the emergence of complex cell-like RFs. In this work, we extend these studies to stereoscopic natural movies and show that by optimizing the stability of visual representations *disparity selective* and *translation invariant* cells emerge. Moreover, the response properties of these cells resemble in many aspects binocular neurons observed in visual cortex. The concept of optimally stable representations successfully describes important aspects of the binocular processing of neurons in primary visual cortex. This is compatible with the view that the synaptic organization in the upper cortical layers may be governed by a simple mathematical rule.

2.2 Introduction

Starting from early work on retinal electrophysiology (Kuffler 1953, Barlow 1953), continuing with the Nobel Prize winning discovery of Hubel and Wiesel (Hubel & Wiesel 1962a) in primary visual cortex, a wealth of data has been made available on the properties of neuronal responses in the visual system up until now (Ringach 2004). We know that neurons in V1, the cortical area closest to retina, respond to visual stimulation with specific characteristics. Edge-like structures with a specific orientation, spatial width and motion direction evoke vigorous activity. Our understanding of the processing in the mammalian visual system is largely based on such descriptive characterization of RF properties.

Given that real-world data is not labeled as edges, surfaces, textures etc., frameworks relying on unsupervised, feed-forward learning of such RF properties are much appreciated. Early on it has been emphasized that the encoding of sensory signals is related to the statistics of the input they are exposed to (Barlow 1953; 1961b, Field 1987). For example, natural images do contain redundancies e.g. a brightness value measured at a given position is correlated with a neighboring position due to the dominance of low spatial frequencies in the natural images. It has been estimated that natural images contain less than 60 % of the entropy of images having the same power spectra (Chandler & Field 2007), meaning that strong redundancies exist. Therefore representations that are much more efficient than pixel representations of images would be possible. It has been proposed that sensory neurons may reduce redundancies in the input signals by efficiently representing the key building blocks of natural images (Barlow 1961a).

Recent theoretical studies provide insights into the relation of neuronal response properties and natural stimulus statistics especially with an emphasis on simple and complex cells. Simple cells, constituting a considerable number of neurons in V1, are optimally activated by stimuli with specific position, orientation and spatial frequency and it has been shown that this is compatible with an optimally sparse representation of natural visual stimuli (Field 1987, Olshausen & Field 1996). Sparse coding implies highly selective representations of the input signals, and consequently, this results in neurons that are silent most of their lifetime in response to various external input but very active when their preferred stimulus is in their RF. Olshausen & Field (1996) created optimally sparse representations of a large set of natural images in a neuronal network simulation. They found that the features that these artificial neurons are selective for share many common characteristics (e.g. localized RFs, orientation and spatial frequency selectivity) with simple cells found in V1.

Complex cells, the other dominant population in V1, are similar in many respects. However an important qualitative difference is that their responses are invariant with respect to the precise position of the stimulus (Hubel & Wiesel 1962b). The principle of temporal coherence (also referred as slow feature analysis) was originally proposed in the form of the trace rule (Földíak 1991). The main idea is based on the assumption that the relevant features in an environment change much slower (e.g. orientation of an edge) than the noise component (e.g. neuronal noise). Therefore by learning to extract slow-changing features from a signal, one might expect to gather a set of useful features. It has been shown that the position invariance property of complex cells can be understood computationally within the framework of forming optimally stable representations of natural movies (Hyvärinen & Hoyer 2001, Einhäuser et al. 2002, Wiskott & Sejnowski 2002, Körding et al. 2004). Optimizing the stability of representations has been also successfully used to predict the non-linearity of neurons (Kayser et al. 2003a). Thus, important characteristics of simple and complex cells in V1 can be understood in terms of general properties of the representations formed.

All these studies have implicitly assumed a monocular visual system as the simulations were realized using natural movies or images recorded with only one camera. This discards therefore the binocular disparity dimension of the visual feature space. This is obviously an acceptable starting point but also a considerable simplification of the visual system of animals. Neurons in the primary visual cortex of monkeys and cats which have two frontally located eyes with a large binocular field, exhibit prominent binocular interactions (Pettigrew et al. 1968, Poggio & Fischer 1977, Anzai et al. 1999a;b, DeAngelis 2000).

In such animals, fixational eye movements bring a given point in a scene to the foveal region, which contains a high density of photoreceptor cells, and because of the convergence of the line of sight to the same external object, both left and right retinal images are in register with respect to each other on this retinal region. However, all other parts of the external world project at disparate positions on both retinae. This is reflected in the interesting differences in the spatial organization between left and right retinal counterparts in the RF of binocular neurons. It is thought that neurons that are selective for these differences are suitable for the detection of disparity (Ohzawa et al. 1990) and underpin the basis of stereoscopic perception (Cumming & DeAngelis 2001). These differences in the RF spatial organization take the form of shifts in the position of the RF center or small phase shifts in the precise location of ON and OFF subregions (Anzai et al. 1997). Moreover it has been reported that shifts of the RFs are specially biased toward a horizontal direction, paralleling the statistical fact that the most of the disparities occur along the horizontal axis due to the alignment of both eyes (Cumming 2002). On the other hand intriguing facts such as the presence of binocularly unbalanced neurons are also observed. Whereas some binocular neurons do have rather balanced input from both eyes, a non-negligible amount receives unbalanced input (Wiesel & Hubel 1974, LeVay et al. 1985). However, the reason for this organization is not well-understood. On the psychophysical level, it is well known that the binocular disparity feature, even on its own, is a powerful cue to create a vivid sensation of depth, indicating the fundamental aspect of this feature detected by these neurons (Julesz 1960). Therefore, a complete account of the RF properties must include the binocular dimension. Moreover, enriching the input by including this visual feature offers us the possibility to test different coding schemes on the basis of how well they reproduce classically observed binocular characteristics.

In this Chapter, we record stereoscopic movies with a pair of micro-cameras carried by a freely moving cat in a natural environment and we train a neuronal network to make optimally stable representations of these stereoscopic natural movies. A comparison with the reported physiological data tests whether we can understand binocular neuronal RFs as optimal representations.

2.3 Methods

2.3.1 Recording of Videos

We used a custom made setup to record stereoscopic movies. Two micro-cameras (The Imaging Source, DFM-5303) were mounted on a crossbar with a distance of 4.8 cm (Fig. 2.1) approximating the interocular position shift of the cat. The setup was calibrated to obtain a binocular parallax of zero at infinity. The gain control of both cameras was fixed to a constant value to avoid luminance differences. In order to synchronize two video streams we used a manually controlled shutter closing circuit that produced simultaneously black images in both cameras. The total weight of the setup was 71 g and was reversibly attached by two screws to an electrophysiological implant of a cat. During recording the cat was freely exploring an outdoor environment and didn't exhibit any observable behavior of dislike to these procedures. The movies were recorded in a forest near the university campus. Care has been taken to avoid any human constructions. Two mobile VCRs (Roadstar, VDR-6205K) carried by the experimenter have been used for the recording. The videotapes have been digitalized in Matlab (Mathworks, Natick, MA) by using a homemade script. With this setup we recorded 21 movies on different days. These correspond to a total of 27556 non-compressed 752×582 RGB movie frames at 25 Hz. The whole procedure is in compliance with institutional national guidelines for experimental animal care.

2.3.2 Building the Stimulus Set

The stimuli to train the network consisted of a sequence of grayscale image patches taken from the stereoscopic movies. Sequences were two frames long and corresponded to an interval of 40 ms due to the sampling rate of the micro-cameras. The patches were 40 by 40 pixels, roughly corresponding to 1 degree of view angle. For computational reasons, we down sampled these patches by half.

A *stereo-patch* was formed by extracting a pair of monocular patches from the same pixel

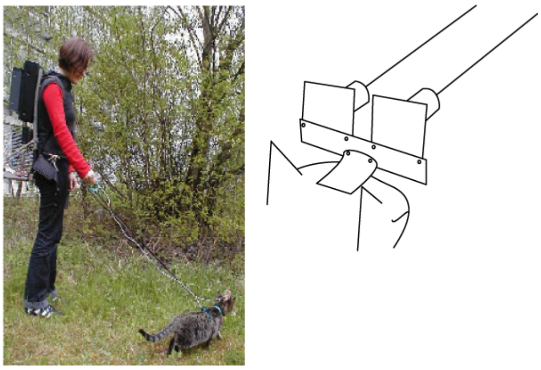

Figure 2.1: **Recording of Natural Movies.** Natural movies were recorded at a sampling rate of 25 Hz using a pair of analogue cameras reversibly mounted on the heads of cats. As the cameras are positioned to have 0 binocular parallax at infinity the only disparities presents in the input are crossed disparities. The cameras were positioned so that the Cameras and plastic support (weight 70 g) were connected by cables to a pair of analogue video recorders carried by the experimenter. Cats did not exhibit any signs of discomfort because of the additional weight on their head, and freely explored their surroundings. See also the Fig. 1.1D

coordinates of corresponding left and right images and by appending them horizontally. This created a rectangular patch of 20 by 40 pixels which was processed as a unit i.e. no explicit distinction on the eye identity (left or right) was provided.

The left and right parts of stereo-patches were subject to different selection criteria based on the similarity between each monocular halves. For the first set we did not apply any selection criteria, therefore the similarity between left and right patches reflected the intrinsic similarity present in binocular images. This set contained stereo-patches composed out of monocular halves that were extracted from randomly selected pixel coordinates (*random condition*). The second set (*confidence condition*) contained stereo-patches composed exclusively out of similar monocular patches. This was intended to mimic the role of the eye movements bringing the binocular images as much as possible into register. The strength of the similarity was evaluated by a confidence measure based on the correlation between left and right patches. We computed the correlation at different horizontal and vertical lags between both monocular patches which had their mean brightness values set to zero. We evaluated how much the peak value of this correlogram stands out by measuring its distance to the background correlation and chose only the most correlated pair of patches. Each set used for training the network contained 20000 stereo-patches for time points t and $t + 1$.

As each pixel represents one dimension in the input layer, we sought to reduce the dimensionality of the stimulus set in order to reduce the computational load. A principal component analysis (PCA) was used to reduce the dimensionality of the stimulus set. PCA was realized on the stereo-patches. The components ranging from 2 to 101 were used; first PCA component resembling the mean luminance was discarded. These were the weights of the PCA components rather than the raw pixel data which were used for the optimization process. The contributions of different principal components were whitened by dividing each eigenvector by its eigenvalue and thus each principal component had an equal contribution. This is thought to be similar to the operation of LGN (Dan et al. 1996) which whitens the contribution of different frequency channels.

2.3.3 Structure of the Neuronal Network

The model implements a two-layer feed-forward neuronal network (Fig. 2.2). The weights connecting the input layer to the subunits in the next layer are subject to modification by the optimization procedure. The activity of the neurons on the top layer is determined by the neuronal model which dictates the way how the activity of subunits converges to the output neuron. This neuronal model is not subject to modification throughout of the simulation and it is similar to the energy model of complex cells proposed by Ohzawa et al. (1990). According to this model shown in equation (2.1), the activity of the output neuron i at time t, $A_i(t)$ is determined by two full-rectifying binocular subunits

$$A_i(t) = \sqrt{(W^i_{1,j} \cdot I(t))^2 + (W^i_{2,j} \cdot I(t))^2} \tag{2.1}$$

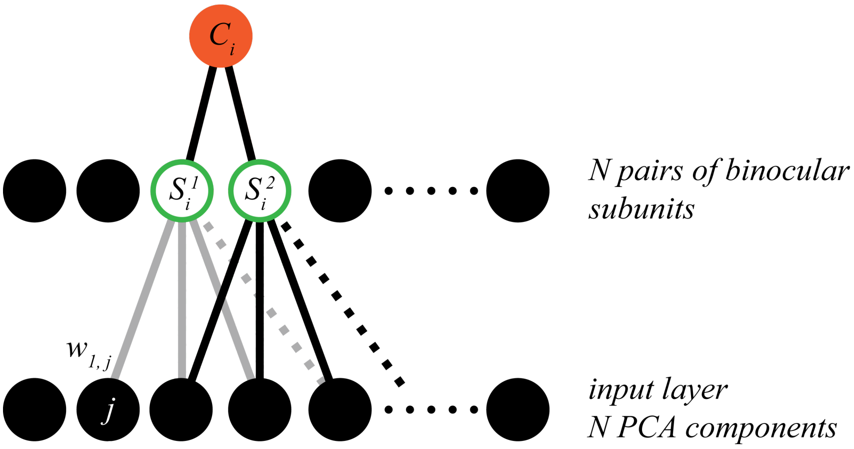

Figure 2.2: **The structure of the Artificial Neuronal Network.** The nodes in the first layer (*bottom* circles) represents input dimensions. The input is formed by temporal sequences of stereo-patches. Rather than using the raw pixel data as the input, the simulations were run on PCA components which covered 90 % of the variance present in the stereo-patches. This was done in order to lower computational load. Each single input dimension were connected to all subunits (*blue* circles). These connections represented the RFs of the subunits. Subunits were organized as pairs and each unit within a pair were connected to a complex cell (*red* circle on top). The activity of the complex cell was determined according to the equation in 2.1. The activity of these neurons were the basis for the optimization process which modified the RFs.

where $A_i(t)$ represents the activity of neuron i at time point t; W^i_1 and W^i_2 represent the RF of the subunit 1 and 2; $I(t)$ represents a stereo-patch at time t. As specified above, no explicit differentiation of the left and right eyes were defined therefore the weight matrix defined by W represents the RF of both eyes. Due to the scalar product preceding the square operation depicted in the equation above, the left and right images interacts linearly for the activation of a given subunit. Thus our network is in accord with the classical energy model Ohzawa et al. (1990). We simulated 100 neurons each with two subunits.

2.3.4 The Goal Function

The activity of the neurons in on the top layer are optimized with respect to an objective function by gradient descent. The goal function is composed of two terms:

$$\Psi_{total} = \Psi_{stable} + \Psi_{decorrelation} \tag{2.2}$$

The first term Ψ_{stable} is the *stability criterion* and defined in equation 2.3.

$$\Psi_{stable} = -\sum_i \frac{< (A_{i,t} - A_{i,t+\delta})^2 >_{stimuli}}{\sigma_{A_{i,t}} \sigma_{A_{i,t+\delta}}} \tag{2.3}$$

The top term is the mean squared derivative of activities with respect to time over all stimuli. This terms is minimized if the activity levels are more stable across time and thus do not change. The brackets symbolize the average over stimuli. δ is set to 40 ms and equal to the time interval between two frames during recording. The trivial solution for this optimization problem is a completely insensitive black RF for both subunits. This would result in a neuron with a constant activity of zero i.e. maximally stable. In order to exclude this trivial solution, we scale the nominator by the total variance in the denominator where σ represents the standard deviation of activity levels of a given neuron at a given time point computed across all stimuli. The resulting expression takes large negative values and thus punishes fast changes. Therefore the optimization is the result of two opposing forces, while the neurons are forced to be as stable as possible across short time scales, they achieve this by being as variable as possible in the long time scales.

Because the stability criterion evaluates individual neurons independently, the optimal solution is identical for all simulated neurons. It is customary to introduce a second, *decorrelation term* to avoid such degeneracy.

$$\Psi_{decorrelation} = -\frac{1}{N-1} \sum_i \sum_{i \neq j} \sigma_{i,j}^2 \tag{2.4}$$

The decorrelation term shown in equation 2.4 forces different neurons to be selective for different features by punishing correlations among neurons. $\sigma_{i,j}$ corresponds to the covariance between neurons i and j. N denotes the total number of neurons. This results in an optimal coverage of the energy landscape at the expense of the stability of individual neuron activity. The overall operation accomplished by the network is to extract the slowest features from a given stimuli set thus achieving a learning supervised by the temporal structure of the stimuli.

2.3.5 Analysis of Receptive Fields

After 800 iterations, consecutive changes of the objective function were smaller than 10^{-5} and the optimization converged. To each monocular RF, a two dimensional Gabor function was fitted. The fit is realized by using the *lsqcurvefit* function of Matlab (Mathworks, Natick, MA). All the fits are graphically checked to ensure that the algorithm did not get caught in a local minima. To precisely compute the phase shift between left and right RFs, we used gratings with optimum orientation and spatial frequency at all possible phase shifts and found the phase shift between left and right monocular RFs.

In order to quantify the translation invariance of the complex neurons, the AC/DC ratio (Dean & Tolhurst 1983) was computed by translating the optimal grating stimulus on the neuron's subunits and computing the activity according to the neuronal model. Based on these activities the ratio in equation 2.5 was computed and assigned to the neuron's AC/DC ratio.

$$AC/DC_i = \frac{\max(A_i) - \min(A_i)}{\text{mean}(A_i)} \tag{2.5}$$

Finally, to compute the binocular dominance index (BDI) the ratio in equation 2.6 was calculated for each neuron. This value deviates from 0 if one of the eyes has a dominant contribution to the neurons activity level.

$$BDI_i = \log(\frac{\max(A_i^L)}{\max(A_i^R)}) \tag{2.6}$$

2.4 Results

We investigate optimally stable representations of the stereoscopic natural stimuli by training a neuronal network. A representative set of the optimally stable RFs are shown in Fig. 2.3. For each of the 6 optimally stable neurons presented in Fig. 2.3, their binocular subunits are presented (left and right columns). The binocular RFs contain ON and OFF regions similar to cortical neurons' RFs. In the following a detailed description and related statistics concerning the selectivities of the optimally stable neurons will be explained for different stimuli conditions. Few non-converged RFs (3 from the *confidence* condition and 8 from the *random* condition) were not included in the analysis.

2.4.1 Orientation and Spatial Frequency Selectivity

The RFs of subunits are selective for a given orientation and spatial frequency and they resemble to Gabor functions (Fig. 2.3). Although in these examples the orientation and the spatial frequency selectivity of different neurons vary a lot, regions pertaining to left and right eye, and first and second subunits of a given neuron are always similar. There is a high correlation between left and right eye orientation selectivity of all subunits ($r = 0.94$ for confidence condition and $r = 0.96$ for random condition). The correlation between same eye RF of different subunits for a given neuron was 0.98 and 0.97 for confidence and random stimuli conditions respectively. Concerning the spatial frequency, the correlation coefficients between left and right RF were 0.99 and 0.98 for confidence and random conditions respectively. Similar correlation exists between the same eyes of different subunits. This indicates a tight coupling of feature selectivity of left and right eyes irrespective of the stimuli set for which the neurons where trained.

The distribution of orientation selectivity over the whole population of neurons has a strong anisotropy towards horizontal and vertical orientations (Fig. 2.4, top row, left panel). Analyzing the distributions of orientations inside patches of 40 by 40 pixels in our natural movie database, we found that the distribution of orientations has a similar distribution to the RF selectivities trained in the confidence condition (data not shown). Concerning the spatial frequency tuning (Fig.2.4, top row center panel) of all neurons, the random condition (gray lines) gives rise to many low frequency selective neurons while in the confidence (black lines) condition neurons cover more evenly the frequency space and extent to higher frequencies. Overall neurons trained by both stimuli sets seem

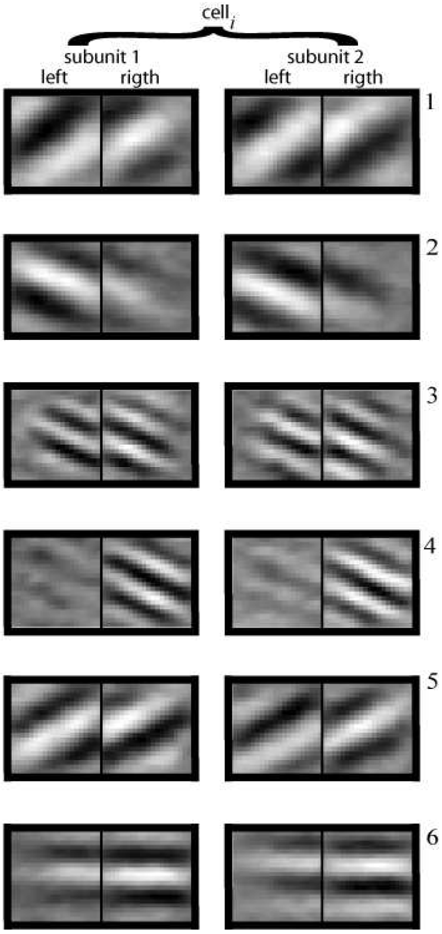

Figure 2.3: **A Selection of RFs Representing Optimally Stable Representations of Stereoscopic Natural Movies.** The binocular RFs of 6 neurons are presented in each row. Each cells is given an index number to specify its location in the next Fig.. Receptive fields of subunits are selective for a given orientation and spatial frequency. A given pair of subunits as well as different eyes of a given subunit exhibit similar preferences. Left and right RFs of subunits of *Cell 1* encode disparities with a shift in the phase of their monocular RFs. On the other side, the *cell 3* encode disparities by positional shifts of the RFs. *Cells 2*, *4* and *6* are binocularly unbalanced. These occur mostly when the cells are trained with less correlated binocular input (*random* condition). Another point worth to mention is that cells seems to be selective for horizontal disparities independent of the orientation selectivity of the cell (see *cell 3* and *6*).

to cover the stimulus space evenly, although this is more prominent in the confidence condition.

2.4.2 Disparity Encoding

Next we investigate the coding of disparity information. In the disparity energy model (Ohzawa et al. 1990), a complex cell can be selective to a given disparity by either shifts in its phase or central position of its left and right RF eye. Both types of mechanisms are thought to be involved in the encoding of binocular disparity in the visual system (Anzai et al. 1997, DeAngelis 2000).

Neurons 1 and *5* in Fig. 2.3 are example cells where the phase shift between left and right RFs of subunits can easily be distinguished. These cells code for the binocular disparity by shifting the phase of their ON and OFF subregions rather than shifting globally their position. The correlation of phase shifts between subunits was 0.98 and 0.96 for confidence and random conditions respectively. This indicates that different subunits of a neuron have very similar phase shifts. The histogram (Fig. 2.4, bottom row left panel) shows that at the population level the subunits cover the phase space mostly between 0 and π which corresponds to the crossed disparities present in the

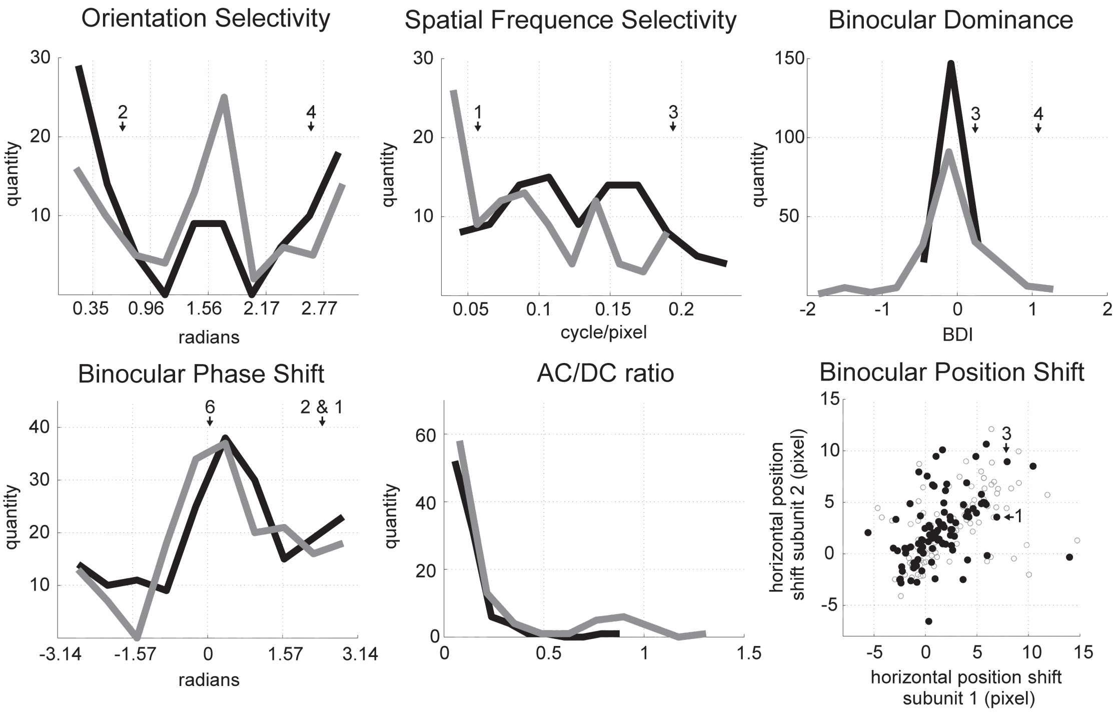

Figure 2.4: **Feature Selectivity of the Simulated Complex Cell Population.** The results of both random (*gray* lines and empty dots) and confidence conditions (*black* lines and dots) are presented as histograms. The numbered arrowheads refer to the indices of neurons depicted in Fig. 2.3.

stimuli.

In addition to encoding disparity by phase shifts, neurons have also horizontal position shifts which make them suitable to encode disparity by position shifts. The *neurons 3* and *6* of Fig. 2.3 are exemplar neurons where the positional shifts are clearly apparent. This can be quantified by analyzing the difference between the centers of RFs of left and right eyes. The center of a RF corresponds to the center of the Gaussian envelope obtained after Gabor fitting process. In Fig. 2.4 (bottom row, right panel), the horizontal position shift between left and right RF of two subunits for a given neuron are plotted against each other (black dots) for the whole population of neurons. It can be seen that position shifts cover a large spectrum of disparities and that the amount of shifts are correlated between each subunits meaning that a given complex cell receiving inputs from a pair of subunits receives similar information with respect to the disparity ($r = 0.5$ and 0.43 for random and confidence condition respectively). This shows that, in addition to phase shift encoding, learnt RFs use also positional offsets to encode disparity as do the RFs observed in physiological experiments.

It has also been reported that neurons in the primate visual cortex, are predominantly selective for horizontal disparities irrespective of the orientation selectivity of the neuron Cumming (2002). The same tendency in the optimized neurons can be observed in Fig. 2.3 (e.g. *neurons 3* and *6*). This shows that the observed cortical specialization for horizontal disparities can be explained by the statistical structure of the binocular visual information in a feed-forward manner.

2.4.3 Translation Invariance

AC/DC ratio quantifies how well a neuron is translation invariant and low values are obtained if neurons exhibit position invariance. The distribution of AC/DC ratios is depicted in Fig. 2.4 (bottom row, middle panel). Most of the values being very close to zero, this shows that neurons are position invariant with respect to the precise localization of their optimal stimuli. The random condition gives generates slightly more neurons having high AC/DC ratios compared to confidence condition. However here also most of the neurons are translation invariant. Taken together with the results presented above, this shows that the optimally stable neurons exhibit, in addition to be selective for a given orientation, spatial frequency and binocular disparity, the property of being invariant to the precise position.

2.4.4 Binocular Dominance

An interesting result is the differences in binocular dominance distribution of RFs trained with different stimulus conditions (Fig. 2.4, top row, right panel). *Neurons 4* and *6* in the Fig. 2.3 are typical binocularly unbalanced cells. These unbalanced RFs were more numerous in the random stimulus condition. This is reflected in a wider distribution of binocular dominance values in Fig. 2.4). This suggests that training with a stimuli where the correlation between left and right monocular images is low, favors the dominance of one eye over the other. We verified this by creating a new set of stimuli where the left and right parts were chosen to be always uncorrelated. This resulted in a complete dominance of an eye over the other in every simulated neuron (data not shown). Diminishing or completely cancelling the contribution of one eye in the face of uncorrelated input seems to be the solution for increasing the stability of the neurons.

2.5 Discussion

2.5.1 Why Do Receptive Fields Have the Structure They Have?

Giving a formal account of the RF structure is one of the most important preoccupations in theoretical neuroscience. While physiologists describe and detect the properties of RFs, theoretical studies attempt to give an account in term of input statistics and some well-defined formal computational constraints such as stability or sparseness. The hypothesis that the process governing RF organization can be well described by an optimization process has received support by experimental and theoretical work (Olshausen & Field 1996, Wiskott & Sejnowski 2002, Hurri & Hyvärinen 2003, Körding et al. 2004). Much of this work concentrated on the selectivity of simple and complex cells in the primary visual cortex with respect to orientation and spatial frequency. However, it is well known that neurons in primary visual cortex are selective to many more features including color, and disparity. The relation between selectivity of neurons for these features and the input statistics is being investigated only recently (Caywood et al. 2004, Einhäuser et al. 2003, Hoyer & Hyvarinen 2000, Wyss et al. 2006). Here in the same direction, we argue that disparity tuning of complex cells can be understood based on the same principles.

We show that the binocular disparity feature can be extracted by forming optimally stable representations of stereoscopic natural movies. Most importantly, the stable representations are in good agreement with the properties of disparity selective simple and complex cells recorded from the cortex (Anzai et al. 1999a;b). We have shown that neurons encode disparities both by binocular phase shift and position shifts similar to the observations noted in the physiological literature (Anzai et al. 1997). Both strategies seem to be the natural outcome of extracting stable features of natural stereoscopic movies. Furthermore, the binocular disparity selectivity seems to be more specialized for horizontal disparities independent of the orientation selectivity of the neurons (Cumming 2002) although a more detailed analysis is needed to quantify this. Additionally the binocularly unbalanced RFs, described in the early physiology literature (Hubel & Wiesel 1962b) is shown to emerge from the convergence of uncorrelated input from both eyes. In summary, constraining artificial neurons to be as stable as possible in response to a stream of appropriate input reproduces many aspects of binocular processing that are well known to physiologists.

Studies investigating unsupervised learning of disparity selective neurons are relatively recent. This is probably because it is technically difficult to record such images and movies. On the other hand, disparity feature offers another dimension in which the investigators can evaluate the similarities of the artificial and biological RFs and evaluate general mathematical principles of synaptic learning leading to the observed RF structures. Hoyer & Hyvarinen (2000) successfully applied a similar approach to stereoscopic images by using independent component analysis. They obtained RFs similar to simple cells resembling in many aspect physiological RF although they did not make a detailed comparison to the physiological RFs.

In the visual cortex of a variety of animals, sophisticated binocular interactions are described. For example, this includes neurons selective for a given slant of the visual stimulus or having selectivity for relative disparity emerging in V2 of monkeys (Thomas et al. 2002). From a theoretical point of view, it is fundamental to know whether the same goal function, when iterated several times, would give rise to RFs comparable to those found in visual areas located hierarchically higher.

As a cell model, we used a classical disparity energy model. Therefore, in each subunit the weighted inputs by left and right eye are summed before the non-linearity is applied. This model has already been used to describe the disparity selectivity of cortical neurons (Ohzawa et al. 1990). But it is also known for its limitations e.g. its inability to capture combinations of inhibitory and excitatory effects from left and right RFs (Read et al. 2002). This choice limits the faithfulness of describing known physiological data. Nevertheless, we considered it more important at the present stage to use the identical type of neuronal model as in previous studies by others and our own group. This decision makes the integration of different feature selectivity, e.g. color (Einhäuser et al. 2003) straightforward. The alternative, to use a different neuronal model to understand disparity, and to worry over other features later, is of course also possible. Thus, we consider the present study as one of two equally valid approaches to a description of multiple feature selectivity using detailed neuronal models.

2.5.2 How Could Stable Features Be Extracted by Neurons?

How could stable visual features might be extracted by neurons? A set of experiments by Frégnac & Shulz (1999) shows that artificially increasing the covariance between the afferent input and postsynaptic activity by injecting intracellular currents is sufficient to induce long-lasting modifications in the visual responses of neurons. This confirms the possibility for Hebbian-like learning processes in the early visual cortex. It is possible to evaluate these results from the point of view of a given visual feature's temporal stability: the more a feature in the stimulus space is stable over time, the more the postsynaptic neuron representing this feature will tend to be consistently activated by the presynaptic cell. This will fulfill the requirements of an Hebbian learning and consequently this feature will be learnt by the network. At the cellular level, back-propagating action potentials (Larkum et al. 1999) are a natural candidate for generating a local learning signal (Körding & König 2001). Recently Sprekeler et al. (2007) provides a detailed account on how stability can be extracted in a biologically plausible manner using the spike timing dependent plasticity rule.

2.5.3 Relation to Sparseness

In their milestone report (Olshausen & Field 1996), Olshausen & Field showed that artificial neurons representing natural images in an optimally sparse way, have RF structures similar to the ones recorded during electrophysiological experiments. Two constraints were imposed during the training of the neuronal network. The first constraint required the distribution of neuronal activities to be as sparse as possible. This was obtained by imposing a non-Gaussian activity distribution with a strong peak and heavy tails. This forced neurons to be as selective as possible in their responses by having zero or very little response to most of the natural images and strong responses to only a small subset thereof. Worth noting is that the stability criterion does not impose any distribution on the resulting activity levels.

The second driving force of the optimization Olshausen & Field (1996) used constrained the neuronal population to convey as much information as possible regarding the stimulus so as to minimize the reconstruction error of the stimulus. However, this latter constraint cannot easily be justified on biological grounds, simply because neurons in the visual cortex that are processing incoming signals do not "know" anything about what is to be represented. They are simply local agents blind to the global aspects of the incoming signals. In our simulations we did not need to provide any explicit knowledge on what has to be represented. This was motivated from the fact that there is *a priori* no necessity from the neuronal perspective to represent an image.

Another major difference between the sparseness or stability criterion as goal functions concerns the nature of the sensory input. Whereas sparseness, as it is generally implemented, operates on static images, extraction of stable features is tightly related to the bodily motion of the animal, which is in turn responsible for the temporal scale of the changes in the retinal input. Therefore the development of RFs under stability criterion is tightly related to the specific action repertoire of a given animal. It is tempting to speculate that animals of different types may use the same synaptic learning principle in order to extract features from the environment that are relevant for their idiosyncratic sensory-motor action.

3

An Optical Imaging Study: Cortical Dynamics During Processing of Natural Movies*

3.1 Context

In the previous section, we have seen that relatively simple mathematical expressions governing synaptic connectivity patterns captured many essential properties of neuronal RFs located in the early sensory areas. Rather than predicting the responses of complex and simple cells to natural movies *per se*, we were dealing with an optimization problem where the responses were forced to satisfy some well-defined constraints. In this Chapter, we will directly deal with the characterization of neuronal population responses recorded from the superficial layers of the early visual cortex of cats. Natural scenes are complex in nature and there are no simple mathematical formulas with which to describe them; however, they are far from random and they contain strong regularities (Chandler & Field 2007). Given that natural images occupy only a small fraction of all possible images, it is expected that neurons constrained by evolutionary, metabolic and computational pressure have developed some adaptational strategies for the specific input they are exposed to (Graham et al. 2009). We will investigate whether or not cortical responses do exhibit any signs of functional specialization for the processing of natural real-world input. As stimulation, we will use natural movies from the previous Chapter in conjunction with simplified laboratory stimuli, which are by far the most common source of stimulation in physiological experiments investigating visual processing. We will use a state-of-the-art neuronal recording method: Voltage-Sensitive Dye Imaging (VSDI), in order to visualize directly the responses of neuronal populations to

---

*This Chapter is the result of a tight and fruitful colloboration with Dr. Dirk Jancke in Ruhr-University Bochum. All the experiments are realized by him and his team in Bochum. The recording of movies, preparation of stimulation protocol and all data analysis are realized by me. The results presented here will form the basis for three different publications. As of today, two of these are submitted (Onat et al. 2009c;b), and the third one (Onat et al. 2009a) is close to the final state.

stimulation with our natural movies. Voltage-Sensitive Dye Imaging offers an unequalized conjoint spatial and temporal resolution (Grinvald & Hildesheim 2004). Whereas it can simultaneously record neurons spread over distances of few millimeters on the superficial layers of the cortex, it offers an extremely good temporal resolution allowing the detection of activity changes in millisecond time-scales. However, in order to avoid complications associated with the stereoscopic presentation of natural movies, here we presented our movies monocularly. This recording method is used here for the first time in conjunction with natural movies.

In the Section 3.3 of this Chapter, we compare the spatio-temporal dynamic activity characteristics of neurons driven by simple laboratory stimuli (i.e. drifting square grating) and complex natural movies under controlled stimulation conditions. Our results demonstrate that the processing of natural movies is realized by a rich repertoire of spatio-temporal activity dynamics which are locked to the motion signals inherent to natural movies (Sections 3.3.3.2 and 3.3.3.3). As expected from the simplified nature of the artificial grating stimuli, the spatio-temporal dynamics were much simpler during processing of grating stimuli. Nevertheless we observed a high degree of spatio-temporal inseparability in the activity levels. Our results presented in the Section 3.3.3.4 show that the responses to gratings were characterized by a linear superposition of an orientation and retinotopical map simultaneously encoding the orientation and the position of the displayed grating.

By comparing large-scale activity levels, we show that the processing of natural movies is characterized by a state that is very different than the grating stimuli (Section 3.3.3.5). This state is mainly characterized by an excess of net inhibition. We show that these differences are unlikely to arise from local differences of luminance contrast levels between these two classes of stimuli, but rather result from an evolutionary functional adaptation of the cortex in response to the input patterns it receives under operational conditions.

The last part of this Chapter (Section 3.4) is devoted to the question of contextual modulation under natural conditions. The early visual cortex contains an extended network of lateral connectivity going far beyond the spatial extent of the classical RFs (Seriès et al. 2003, Gilbert & Wiesel 1989). To study the functional role of these long-range lateral connections under natural conditions, we presented our natural movies locally through either one or two Gaussian apertures. Our results presented in the Section 3.4 demonstrated that the local processing of natural movies is influenced by the distant stimulus located many visual degrees away. This contextual stimulation had a net facilitatory effect on activity levels (Section 3.4.3.3) and shortened the latency of the cortical responses (Section 3.4.3.2).

3.2 Analysis of Optical Imaging Data

In this section I provide a detailed account of the general analysis methods used throughout this Chapter. In order to not disturb the flow of thought related to the scientific questions with highly technical issues, I require the reader to refer to the Appendix (in 7.1) at the end of this thesis.

3.3 Are Simple Laboratory Stimuli Processed in the Same Way as Natural Input?

3.3.1 Introduction

The early visual cortex comprises an extended, densely interwoven network, acting on millisecond time scales (Callaway 1998). Across cortical layers, activity is rapidly distributed by local feedback loops (Tucker & Katz 2003, Douglas & Martin 2007), tangentially, long horizontal fibers connect distant neuronal populations to each other (Gilbert & Wiesel 1989, Kisvarday et al. 1994, Grinvald et al. 1994, Bringuier et al. 1999, Jancke et al. 2004). Additionally, feedback loops involving higher areas add to interactions over large distances (Bullier et al. 2001), in sum forming a repeating scheme of connectivity that allows far-reaching, spatio-temporal integration of information. Furthermore, neuronal activity in early visual cortex covaries with basic stimulus parameters such as position, orientation, direction of motion, color, spatial frequency, and binocular disparity. In many mammals, the large scale organization of early sensory cortices is in the form of overlaid maps in which neurons with similar tuning properties are spatially clustered (Hubel & Wiesel 1974, Blasdel & Salama 1986, Bonhoeffer & Grinvald 1991, Hubener et al. 1997, Coppola et al. 1998). When stimulation parameters are varied in isolation or in well-controlled low-dimensional steady-state experimental settings (Jancke 2000, Geisler et al. 2007, Benucci et al. 2007), the layout of cortical maps changes (Basole et al. 2003), but is still predictable using feedforward models of visual processing (Mante et al. 2005).

As expected from the success of these models in predicting changes in population activity and related changes in map layouts, individual neuronal responses stimulated with simple stimuli such as oriented bars and gratings are well predicted by linear models incorporating non-linearities at the output stages (Chichilnisky 2001, Touryan et al. 2002, Rust et al. 2005). However, the success of the prediction is largely contingent on the precise properties of the stimuli that are used; performance drops considerably when non-parametric, complex stimuli, which appear in ecologically relevant settings, are used to drive cortical neurons (Smyth et al. 2003, David et al. 2004, David & Gallant 2005). The statistical structures of natural images are by and large strikingly different to simple laboratory stimulation. While the latter class of stimuli is controlled precisely by the experimenter and thus tailored for the purpose of the experiment, the former does not allow parametrical modifications but better represents the ecologically relevant input to the sensory system. Despite their complexity, natural images and movies do possess intrinsic regularities and recent research has shown that cortical (Felsen et al. 2005b, Mante & Carandini 2005) and subcortical (Dan et al. 1996) neuronal machinery incorporate diverse phenomena adaptive to the statistical structures of the input signals. While the question of how individual neurons are adapted to the input statistics has attracted a great deal of effort, only a very limited number of studies have compared the large-scale effects of natural and artificial stimulation, and how real-world input is represented cortex-wide within the early sensory areas remains largely unknown.

Simultaneous recording of the large number of neurons involved in the processing of dynamic natural stimuli may in fact lead to an understanding of discrepancies in the results obtained with different classes of stimuli. Here we used a large-scale recording method to investigate neuronal activity dynamics. Optical imaging using blue voltage-sensitive dye records the sum of synaptic currents with high a spatial and temporal resolution across cortical distances involving large numbers of neurons with an emphasis on supragranular layers (Grinvald et al. 1984, Arieli et al. 1996, Shoham et al. 1999, Petersen et al. 2003a, Jancke et al. 2004, Sharon et al. 2007, Ferezou et al. 2007). It therefore reduces the possibility of having a biased sample of neurons and faithfully provides the state of cortex under different stimulation conditions.

Our main working hypothesis is that the cortical circuitry does not embody a fixed, generic processing structure in which all input signals are processed with the same processing characteristics. In particular, the balance between inhibitory and excitatory neurons, being subject to modification (David et al. 2004), may allow different processing regimes for different categories of stimuli. Thus the processing of different stimuli may in principle be realized in different operating regimes, defined by the general characteristics of the state of a large population of neurons as they process their input. One corollary of these considerations is that simple laboratory stimuli tailored for experimental purposes, such as bars in the visual domain and sines or pure tones in the auditory domain may indeed lead the system into a processing regime that is different than the one within which naturalistic stimuli are processed. We defined here an operating regime as the first, second and higher order properties of activity levels derived from large number of neurons. As optical imaging with voltage-sensitive dyes quantifies the net difference between excitatory and inhibitory sources of synaptic activity, it is a well-suited recording method for evaluating differences that characterizes different operating regimes.

In order to shed light onto the cortex-wide large-scale processing of simple and natural visual input we presented natural movies and compared a variety of response characteristics to the activity levels evoked by simple laboratory stimuli. Our results show that the processing of natural movies within the superficial layers of early visual cortex occurs under a different spatio-temporal context and in a distinct operating regime.

3.3.2 Experimental Procedures

3.3.2.1 Stimulus Acquisition and Presentation

Natural movies were recorded by freely moving cats with a head-mounted pair of cameras (DFM-5303, The Imaging Source Europe, Bremen, Germany) while exploring a forest (Fig. 1.1D and 2.1). In this study we used the data recorded from one of the cameras only, hence no binocular depth cues were given. The cameras and the supporting frame (≈70 gr) were reversibly attached to an electrophysiological implant.

The gain of the camera was set to a constant value and the frames were recorded with a sampling rate of 25 Hz. The output of the camera was recorded with a VCR (Roadstar, VDR-6205K, Novazzano, Switzerland) carried by the experimenter. Recorded analogue movies were later digitalized and transformed to grayscale with at a size of 640×480 pixels. The average and RMS value of brightness was then equalized to the respective average values of the movies in our database so that each frame in our database had globally the same average luminance and RMS value. We excluded movies that were solely characterized by immobile fixation behavior of the cats.

A 24” Sony monitor (Sony Triniton GDM-FW900, Tokyo, Japan) covering a visual field of $30° \times 40°$ was used to present the stimuli with refresh rate of 100 Hz (thus each cat-cam movie frame was repeated 4 times). The stimuli were shown using Matlab (Mathworks, Natick, MA, USA) and Psychophysics Toolbox extensions (Brainard 1997, Pelli 1997). The duration of the presentation was 2 seconds including 200 ms prestimulus period. Stimuli were presented binocularly on the monitor located 50 cm distant from the cat's eyes. Eyes were converged using a prism in front of the eye that was ipsilateral to the recorded hemisphere. The mean achromatic luminance of the stimuli was 11 cd/m^2 with 54 cd/m^2 for brightest, and 1.6 cd/m^2 for darkest pixels. The average brightness of the screen during stimulation and interstimulus intervals (15 s) was kept constant during the course of the experiment. Stimuli were presented in pseudo-random order in blocks.

3.3.2.2 Stimulus Conditions

During full-field conditions natural movies and gratings were shown across the entire monitor screen. In each experiment, two different natural movies (movie 1 and movie 2) and gratings (horizontal and vertical) were used. Square wave gratings with spatial frequency of 0.2 cycles/° and drifting velocity of 6 Hz were shown in horizontal and vertical orientation. Their mean luminance and RMS contrast were adjusted to match the natural movies.

3.3.2.3 Experimental Setup

All surgical and experimental procedures were approved by the German Animal Care and Use Committee (AZ 9.93.2.10.32.07.032) in accordance with the Deutsche Tierschutzgesetz and the NIH guidelines. In brief, animals were initially anesthetized with Ketamine (15 mg/kg i.m.) and Xylazine (1 mg/kg i.m.), supplemented with Atropine (0.05 mg/kg i.m.). After tracheotomy, animals were artificially respirated, continuously anaesthetized with 0.8-1.5% Isoflurane in a 1:1 mixture of O_2/N_2O, and fed intravenously. Heart rate, intratracheal pressure, expired CO_2, body temperature, and EEG were monitored during the entire experiment. The skull was opened above A18 and the dura was resected. Paralysis was induced and maintained by Alloferin®. Eyes were covered with zero-power contact lenses as protectives. External lenses were used to focus the eyes on the screen. To control for eye drift, the position of the area centralis and RF positions were repeatedly measured. A stainless steel chamber was mounted and the cortex was stained for 2-3 hours with voltage-sensitive dye (RH-1691), and subsequently washed out with artificial CSF .

3.3.2.4 Retinotopical Mappings

Prior to optical recordings, the topographic mapping between the cortical surface and the visual field were scrutinized by means of several electrode penetrations (Fig. 3.1; penetration sites and corresponding RFs are color coded). Stimuli were positioned so that the upper movie patch matched the RF position of the simultaneously recorded multiunit activity (Fig. 3.1, red circle and rectangle). This ensured that the distance between the Gaussian masks were separated beyond the size of classical RFs.

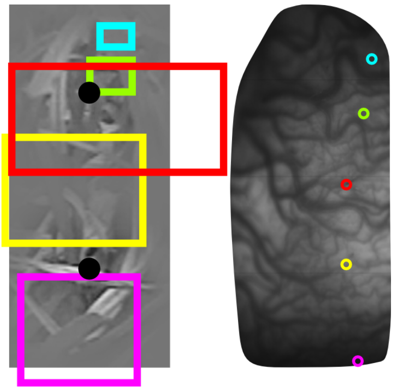

Figure 3.1: **Retinotopical Mapping.** Retinotopy across the imaged cortical region was evaluated by hand mapping of RF locations (colored rectangles) at various penetration sites (see corresponding colored dots in vascular image). Parallel to image acquisition, multiunit activity was recorded at a position (*red* dot) in which the neurons' RF approximately fitted the center of the upper stimulus (*red* rectangle; *black* dots mark center of each movie patch).

3.3.2.5 Data Acquisition

Optical imaging was accomplished using an Imager 3001 (Optical Imaging Inc, Mountainside, NY) and a tandem lens macroscope (Ratzlaff & Grinvald, 1991), 85 mm/1.2 toward camera and 50 mm/1.2 toward subject, attached to a CCD camera (DalStar, Dalsa, Colorado Springs). The camera was focused ≈400 nm below cortical surface. For detection of changes in fluorescence the cortex was illuminated with light of 630±10 nm wavelength and emitted light was high-pass filtered with cutoff at 665 nm using a dichroic filter system. Cortical images were acquired at a frame rate of 220 Hz covering regions of approximately 10×5 mm of A18.

Our experimental data that we base our analysis contains 11 hemispheres (10 animals) for the analysis reported in the Section 3.3. For electrophysiological recordings before and in parallel with VSDI, an in house-build device was used that allowed targeted penetrations at different locations without opening the sealed recording chamber.

3.3.2.6 Preprocessing

The raw data was processed in two steps. First, in order to remove differences in illumination across different pixels, divisive normalization is performed on all the recorded raw samples of a given pixel by its DC level during prestimulus period. Thus the average value of each pixel during this period is exactly equal to one for all pixels. Divisive normalization is preferred over subtractive normalization as it also equalizes for standard deviation of different pixels. This is due to the linear relationship between the prestimulus DC amplitude and standard deviation of the pixels' activities. Second, the removal of heart-beat and respiration related artifacts were realized by subtracting the average blank signal recorded in the absence of stimulation. The differences are later normalized by the blank signal in order to be independent of the global activity level fluctuations which occurs during the course of an experiment. As our recordings were synchronized with the heart-beat cycle of the animal, this blank subtraction step effectively removes these artifacts. Moreover this method is preferred over the cocktail blank correction because our conditions were not composed of orthogonal stimuli. These steps were applied for each trial separately and the outcome was averaged across trials. The number of trials ranged from 20 to 37 for different experiments.

3.3.2.7 Region of Interest Selection

In cases where we restricted the analysis to a limited ROI of the recorded area, selection heuristics were based on either time averages or peak activity levels. From these average or peak activity maps, ROIs were created according to the percentile ranges of the pixel's activities. For example a given set of pixels could be assigned to belong to the percentile interval of 95-100%, 90-95% and so on. This method of ROI selection has many advantages. As the selection merely depends on the ranking of the pixels based on their activity levels and not their absolute value, it facilitates the usage of the same heuristics for different experiments with varying absolute activity levels. A further benefit is that different ROIs originating from different experiments are all composed of same number of pixels. For comparison of evoked activity levels during localized presentation of natural movies, we focused on the highest 5th percentile of the pixels. However the same results were obtained also with a selection based on the 10^{th} percentile.

3.3.2.8 Latency Detection

Evaluation of the response times was realized on z-score transformed data. Z-score directly translates a given activity level to its distance from baseline in units of the baseline process' standard deviation. It is therefore invariant to changes on absolute level of activities which occurs during different experiments. Baseline activity was characterized by temporarily averaging activity levels during prestimulus period. Standard deviation of the baseline activity for a given pixel was computed by pooling all its prestimulus samples over all conditions; this increased the sample size and therefore reduced the uncertainty of the approximation. Because of the slight spatial dependencies of average and standard deviation statistics, we computed each pixels' average and standard deviations for the baseline period separately. After z-score transformation, latency was determined by detecting for each pixel the time point at which activity exceeded 2 times the prestimulus time standard deviation in two consecutive time frames.

3.3.2.9 Singular Value Decomposition

We applied SVD to the evoked activity by using *svd* command of Matlab software (Mathworks, Natick, MA, USA) after transforming each time frame into a vector. Singular values increased linearly in the logarithmic scale and significant components were detected by identifying the first component which had a significantly different increase in its weight with respect to the previous components.

3.3.2.10 Detection of Confidence Intervals

Due to somewhat large variability across experiments, in some cases it was preferable to compare first natural and grating conditions in a pair-wise fashion within each experiment and then evaluate the significance of the deviations across all experiments. In order to make the pair-wise comparison possible we first averaged across both natural movie and grating conditions. In these cases the statistical tests were carried on 11 observations. In all other cases, all observations corresponding to a given condition across all experiments were used for statistical tests. This amounted to 22 observations per condition.

In order to estimate the confidence intervals we used the statistical bootstrap method. We derived the distribution of the statistic of interest by resampling the observations with replacement 10^5 times. The confidence intervals are given in the text within square brackets following the mean values.

3.3.3 Results

We presented natural movies along with drifting square gratings (horizontal and vertical) to anesthetized cats (11 hemispheres, 10 cats) and performed voltage-sensitive dye imaging on cortical area 18 of cat's visual cortex. Over the past decades during electrophysiological and imaging experiments this model sensory cortex has been extensively investigated with simple laboratory stimuli. Our natural movies were captured by cats freely exploring a natural habitat and contain a rich spatio-temporal structure. They depict natural scenes from the cat's point of view and incorporate temporal dynamics mainly due to head and body motion. In order to improve the generality of our results we presented a total of 9 different movies and in cases where the same movies were used across different experiments, different portions of the movies were used to drive the recorded cortical area. Prior to and during the experiment, the recordings were complemented by retinotopical mappings in order to certify which local parts of the movies were directly received by the cortical tissue that was being imaged.

3.3.3.1 Spatio-Temporal Population Dynamics Evoked by Natural and Grating Stimuli

Both kinds of stimuli were presented subtending a large part of the visual field ($\approx 40° \times 30°$) under equal global luminance and luminance contrast conditions (Fig. 3.2). However due to the intrinsic properties of natural movies, local average luminance and luminance contrast values exhibit some degree of variation across the presentation time leading to discrepancies between global and local statistics. For example, a relatively uniform surface of a piece of a wood or a leaf that is represented with a very peaky distribution of local brightness values would necessarily have a low luminance contrast. As these areas in the image cannot be constrained to have a higher luminance contrast without impairing the quality of the movie we did not control for these local variations, thus local contrast and average luminance in a given restricted area of the movie was occasionally higher or lower than the global target values. Within these restricted regions, the distribution of the local luminance contrast values were slightly negatively ($p < 0.05$) skewed (skewness = -0.18) with a standard deviation of 9.7% (9.3%/10.1%[†], $p = 0.05$) and an median luminance contrast sensibly but significantly higher than the luminance contrast of the grating stimuli (101.25%, Wilcoxon-Signed Rank, $p < 0.05$, see Fig. 3.8 for the distribution of contrast values of all the frames). There was no systematic bias that caused the part of the movies that were directly received by the recorded area to be lower than the global statistics. Therefore we consider that the local fluctuations do not

---

[†] In the following, bootstrap confidence intervals are presented in brackets according to the following convention (Lower Bound/ Upper Bound, p = Confidence Level)

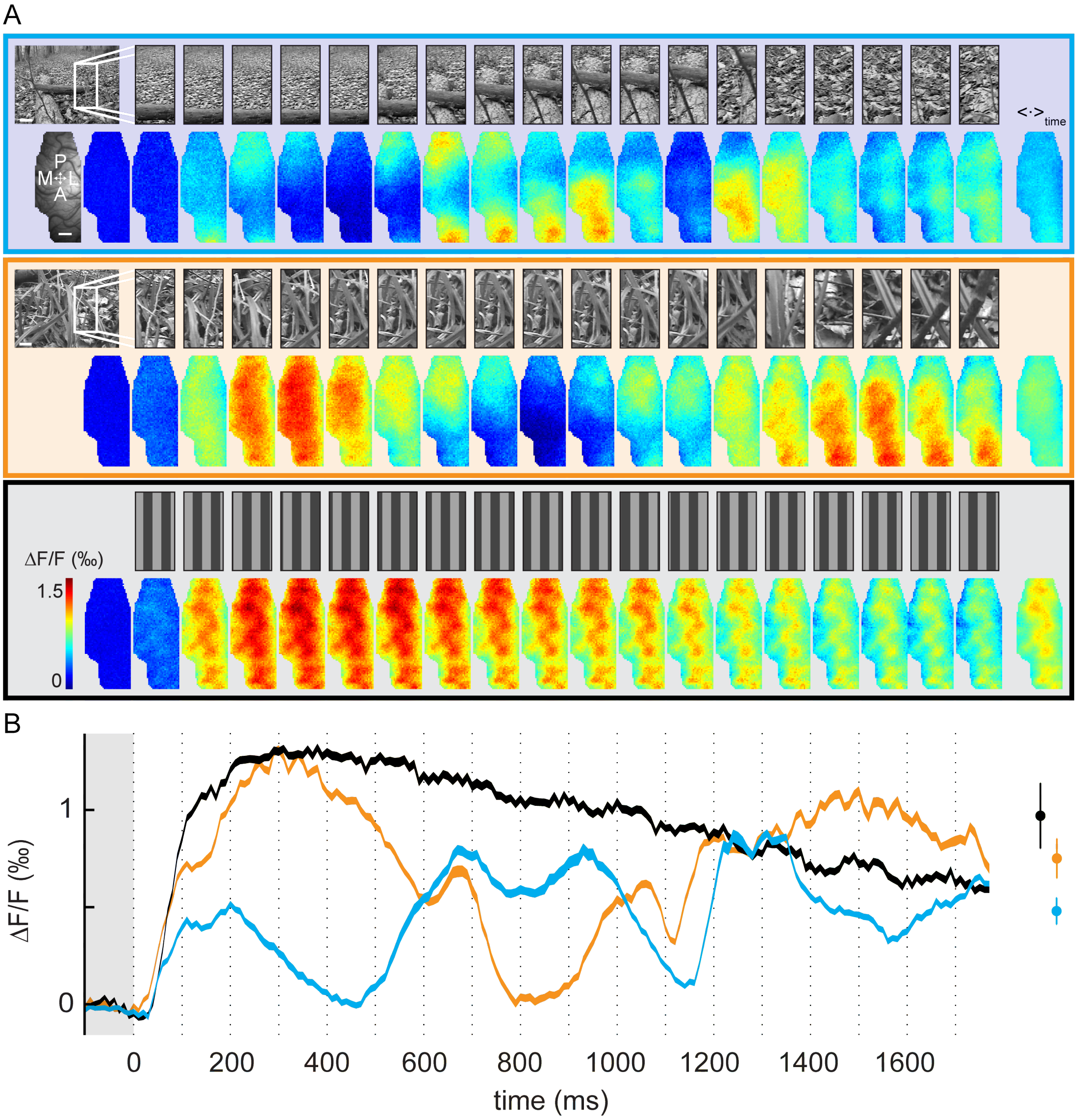

Figure 3.2: **Recording Cortical Responses to Natural Stimuli and Gratings.** **(A)** Two natural movies (within the *blue* and *red* boxes) and vertical gratings (within the *gray* box) used as stimulation are depicted together with evoked optical responses. Visual stimuli are depicted in the upper row within each box. Leftmost image represents an example movie frame covering approximately a visual angle of ≈ 40° × 30°. The scale bar represents 5° of angle of view. The local portion that directly stimulates the recorded cortical region is delineated with a white rectangle. The temporal evolution of the movie within the delineated region is shown as a succession of frames. The second row within each box is space-time representation of the evoked optical signals recorded during two seconds including the prestimulus period. Each frame represents the average activity during intervals of non-overlapping 100 ms. The rightmost image shows the average activity computed over the entire stimulus presentation. Vascular image of the recorded cortical area is shown in the top leftmost frame (P=posterior, A=anterior, M=medial, L=lateral; scale bar represents 1 mm). Colorbar indicates (bottom box) activity levels as fractional fluorescence change relative to blank. **(B)** Time courses of global activity computed by taking the average across all pixels of a given frame. Shaded *gray* area symbolizes the prestimulus period. Line colors are matched to the boxes shown in (A); *black* = grating, *blue/red* = natural conditions. The thickness of lines represents confidence intervals computed by resampling all the pixels that belong of a given frame ($p = 10^{-5}$). Right panel: Amplitudes of activity averaged over the presentation time. Error bars represent the standard deviation of activity levels depicted in the time-course.

constitute a major obstacle for the comparison of activities across stimulus categories and the differences in the activity levels are not straightforwardly attributable to the local differences in stimulus power.

The region which was recorded covered a surface of approximately 1 cm and it was driven by a portion of 3° to 4° of visual stimulus (Fig. 3.2A, white rectangle in leftmost column). It contained a large number of neuronal components spanning multiple orientation columns and pinwheels; it thus constituted a representative sample of an early visual cortex. Fig. 3.2A depicts the evoked activity recorded under different stimulation conditions along with the local portions of the movie frames directly received by the recorded area; the cortical response is represented with 20 frames each representing the average activity during non-overlapping intervals of 100 ms. With the onset of the natural movies an excitatory response was observed within the first 100 ms. (Fig. 3.2A, first and second rows, see also Fig. S1). The average latency computed from the spatially averaged data (Fig. 3.2B) was 57 ms (SD = 12 ms) and it was not significantly different than the latency of 55 ms (SD = 11ms, paired t(21) = -0.51, p = 0.61) observed in response to grating stimuli. However, computing the latency separately for each pixel, we found that not all the pixels responded in the same time to the stimulus onset. The spatial distribution of latencies was much more variable in response to natural movies. To evaluate the spatial inhomogeneity we computed the standard deviation of pixel latencies; the median standard deviation computed over all experiments (n = 11) was 17 ms and 40 ms ($p < 0.001$, signtest) for grating and natural movie stimuli respectively. This suggests that the onset of natural movies do generate a spatially patterned and inhomogeneous activity across the cortical surface.

Following the onset of grating stimuli evoked responses appeared in the form of ripple-like spatial patterns characteristic of the cortical columnar architecture for orientation selectivity (Fig. 3.2A, third row and black trace). This was accompanied by a strong orientation unspecific excitation distributed widely across the imaged area. Activity increased sharply until 300 ms, and then constantly decayed (Fig. 3.2B, black curve). On the other hand responses to natural movies (Fig. 3.2A, first and second rows and red/blue traces), exhibited dynamic spatial modulations characterized by the emergence and propagation of spatial activity patterns. These spatial patterns co-occurred with temporal fluctuations in the overall activity levels. The appearance of large regions of low (e.g. frame 6 in movie 1 and frame 10 in movie 2) activity suggests that excitatory drives were at times either absent or completely counter-balanced by inhibition. At these intervals the average activity within the imaged area approached to prestimulus baseline levels suggesting that the balance between excitation and inhibition were constantly subject to modification.

3.3.3.2 Locking of Activity to the Stimulus Motion

Temporal dynamics of our natural movies were dominated by the body and head motion of the cat. In parallel, we observed that the time-course of the global activity i.e. average activity of the recorded cortical area, during stimulation with natural movies displayed considerable fluctuations. For each movie within the relevant local portion we computed the amplitude of the flow-field vector representing the motion between consecutive frames (Fig. 3.3A black traces; see also Fig. S1). Owing to relatively long

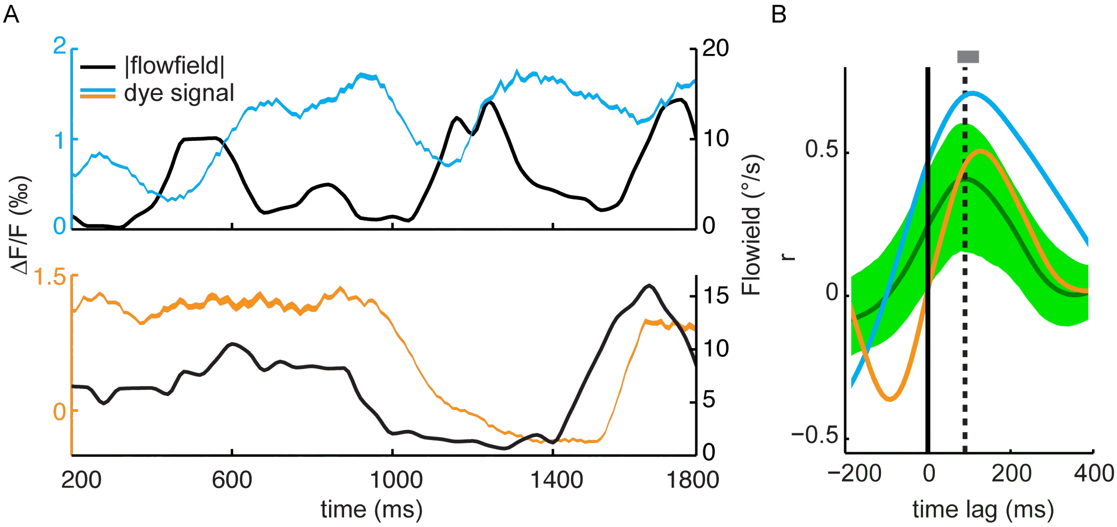

Figure 3.3: **Motion Locking of Average Activity.** **(A)** The temporal progression of spatially averaged activity levels is depicted (*red/blue* traces, left y-axis) together with the absolute amplitude of the motion vector (*black* lines) computed between consecutive frames evaluated within the local portion of movies directly stimulating the recorded cortical tissue. In order to discard the transient part of the response only the period of 200 to 1800 ms after stimulus onset is used. The thickness of lines represents bootstrap confidence intervals ($p = 10^{-5}$) computed by resampling all the pixels that belong to given frame. **(B)** The correlation between the absolute flow field and the time-course of activity computed at different lags of time. The two examples in (A) are depicted with *red/blue* traces. *Green* trace represents the mean taken over all experiments and movies (11 experiments with 2 movies each). *Shaded* area represents 95% bootstrap confidence intervals. The *dotted* line is drawn at 90 ms where the correlation peaks. The *shaded gray* horizontal bar represents the 95% bootstrap confidence interval of the peak location.

recording interval, we quantified the extent to which the observed fluctuations in global activity were related to the absolute velocity profiles of the movies by computing the correlation coefficient at different lags of time (Fig. 3.3B, right panel). For the two examples shown in Fig. 3.3A, the cross-correlogram peaked at approximately 100 ms and reached values up to 0.7 (Fig. 3.3B, blue and red traces). The correlation curve averaged across experiments peaked at 90 ms with a correlation value of 0.41 (0.6/0.15, $p = 0.05$) and was significantly different than zero at the peak position (Wilcoxon-Signed Rank, $p < 0.001$; Fig. 3.3B, green trace). In order to evaluate the variability of the peak positions across different experiments and movies we detected the peak position of the correlation curves between the interval of 0 to 300 ms. The average peak was located at 91.3 ms (71/122 ms, $p = 0.05$, Fig. 3.3B, gray horizontal bar). These results show that motion cues within dynamic natural stimuli entail spatially widespread activity fluctuations by simultaneously driving large numbers of neurons and that a considerable part of the variance present in the global activity time-courses is explained by the velocity profile of natural movies.

3.3.3.3 Spatio-Temporal Separability of Activity Dynamics

We used SVD analysis in order to evaluate the separability of the spatio-temporal dynamics accompanying the temporal modulations reported above. A function $f(x, t)$ describing spatio-temporal dynamics is said to be separable if it can be decomposed into

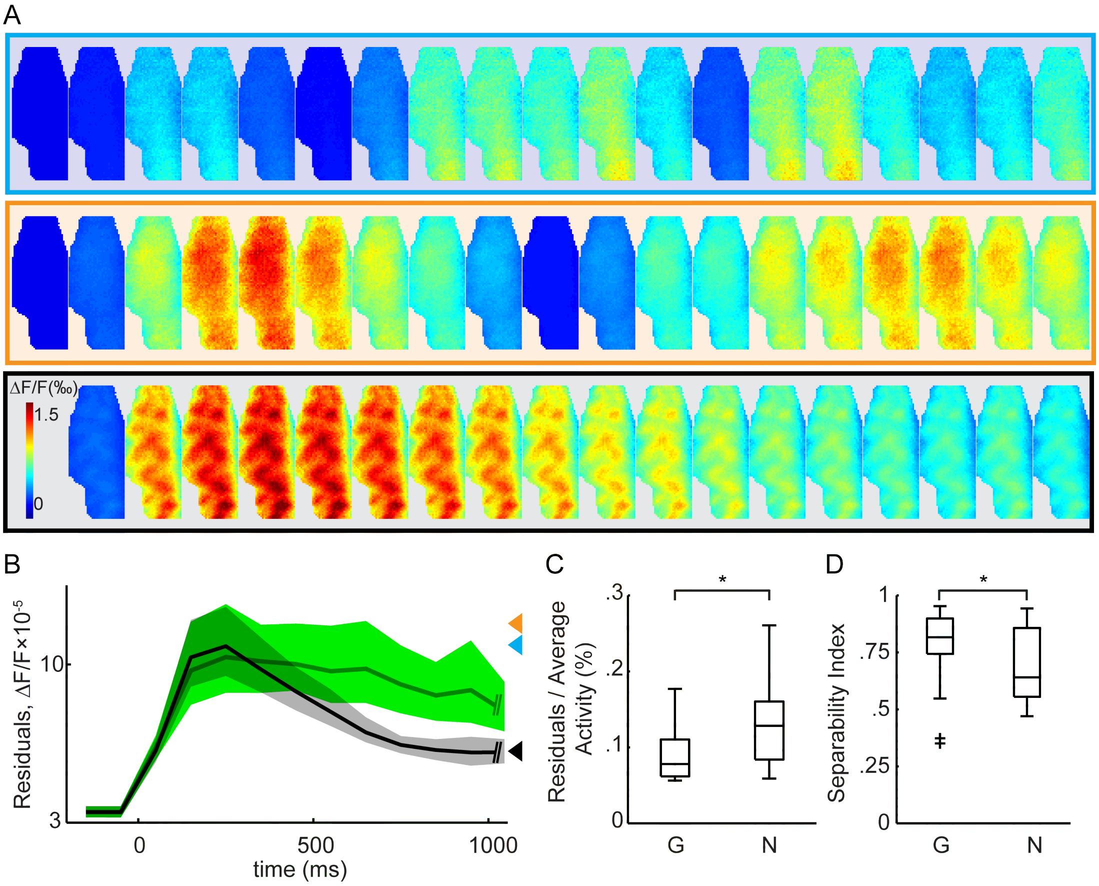

Figure 3.4: **Separability of Spatio-Temporal Activity Dynamics.** **(A)** Reconstructed evoked activity under spatio-temporal separability hypothesis for the same data shown in Fig. 3.2. The data represents the first SVD component and thus the outer product of spatial and temporal activity profiles computed by averaging along temporal and spatial dimensions, respectively. **(B)** Mean time course of the spatially averaged prediction error computed between the reconstructed and recorded evoked activity is depicted for natural movie (*green*) and drifting grating (*black*) conditions. The triangles represent overall average errors for the examples used in (A). Shaded area represents 95% bootstrap confidence interval computed by resampling the data of different experiments and movies. **(C-D)** Distribution of two different measures quantifying spatio-temporal separability is shown as boxplots. The difference between grating and natural conditions are significant when evaluated in a pair-wise manner ($p<10^{-5}$ and $p<0.01$ for **(C)** and **(D)** respectively).

the outer product of its single dimensional constituents, i.e. $< f(x) >_t$ and $< f(t) >_x$, which represent signals averaged across the temporal or spatial dimension, respectively. Thus, under separability hypothesis the effects of spatial and temporal dynamics are assumed to interact independently.

SVD transforms the original signal $f(x,t)$ into a weighted sum of separable functions $\sum_i \gamma_i g_i(x,t)$ where γ_i represents the weight of the component i in decreasing order of magnitudes. Following SVD transformation the first component, $g_1(x,t)$, represents the outer product of the average spatial and temporal profiles. If the signal is separable, this results in all singular values but the first one to be equal to null. Hence the first components totally capture the dynamics of the recorded data.

The signal reconstructed under separability hypothesis (i.e. the first SVD component), for the data shown previously (in Fig. 3.2) is depicted in Fig. 3.4A. In order to get a measure of the separability of spatio-temporal dynamics, we computed the residuals between the predicted activity and the evoked activity; where larger residuals would be an indication of less separable activity dynamics. The time-course of the prediction error averaged across all experiments is shown in Fig. 3.4B (see also color matched triangles). During stimulation with natural movies, the residuals (Fig. 3.4B, green traces) reached a plateau shortly after the stimulus onset and stayed relatively constant thereafter. The error profile under grating stimulation (black curve) started with a phasic strong peak and quickly dropped toward values lower than the observed for natural stimulation leading to overall smaller error values (Fig. 3.4C). The median amplitude of prediction errors was lower in the grating conditions and corresponded to less than 8.8 % (7.72/10.5, $p = 0.05$) of the average activity (Fig. 3.4C) this was significantly smaller than the natural conditions ($p < 10^{-5}$, pair-wise bootstrap) which had a mean error of 13 % (11.1/15.4 %, $p = 0.05$).

The reason for the strong initial peak in residuals can be understood by comparing the initial frames of the reconstructed (Fig. 3.4A, third row) and evoked data (Fig. 3.2A, third row); one can see that the spatial patterning due to cortical columnar organization was much weaker in the recorded evoked activity, rather a widespread unspecific activity at frames between 300 and 500 ms was clearly evident. The separable prediction overestimated the fine-grained columnar structure observed shortly after stimulus onset and thus led to higher prediction errors. This change in the overall spatial pattern was the main source of non-separability during grating conditions.

We also computed the separability index, a previously used measure (Escabi et al. 2003) which quantifies the ratio of the first singular value to the sum of all singular values; it thus represents the variance accounted by the first SVD component and is equal to 1 in case of a separable function. Median separability index was 0.81 and 0.64 for gratings and naturals respectively ($p < 0.01$, pair-wise bootstrap). Accordingly we found that the number of significant singular values necessary to describe the activity during grating conditions were much smaller. While in average 5 (4.3/5.6, $p = 0.01$) singular values were found to be significant, this number was equal to 7.3 (6.4/8.5, $p = 0.01$) in the case of naturals. We thus conclude that activity within superficial layers is characterized by non-separable spatio-temporal dynamics during both natural and grating conditions.

3.3.3.4 Singular Value Decomposition Unravels Two Overlapping Maps.

The result of the SVD analysis applied to another dataset is shown in Fig. 3.5A (upper panel) where the responses evoked by two orthogonal drifting gratings are depicted together with the time-course of the activity (Fig. 3.5A bottom panel) which represents the spatially averaged response. The SVD was applied to the later part of the evoked responses (Fig. 3.5, shaded area) excluding the fast transient onset. For the data presented, we observed 6 different significant singular values (3.5b, top-left), suggesting inseparable complex spatio-temporal activity dynamics in the responses to the moving gratings.

However, the existence of numerous significant SVD components calls into question the functional relevance of the dynamics described by these components, as the true underlying biological process is unknown. Therefore, one cannot directly attribute a biological

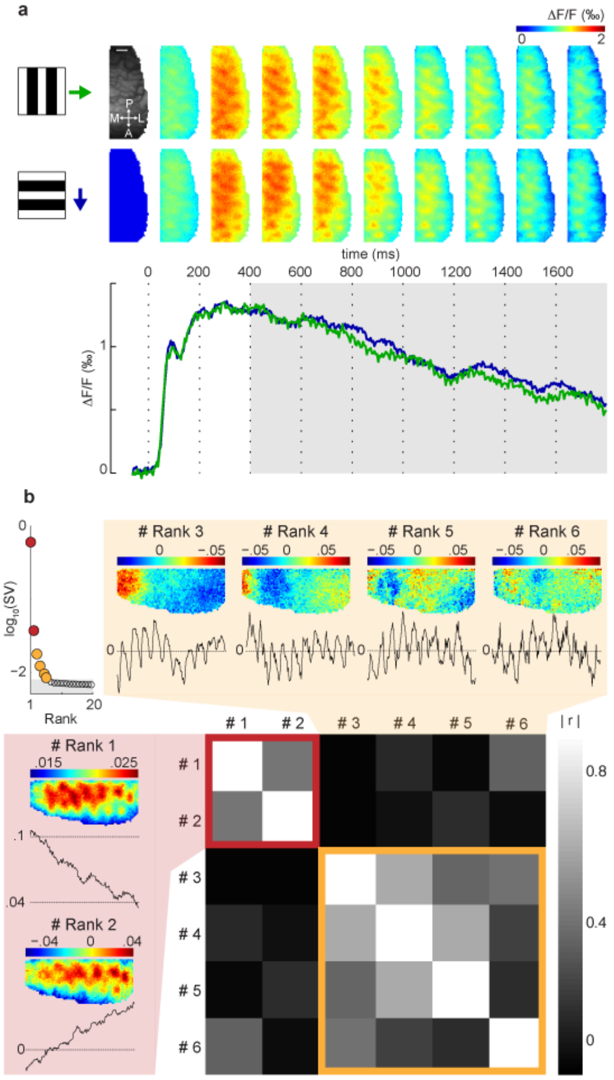

Figure 3.5: **Decomposition of Evoked Cortical Voltage-Sensitive Dye Responses to Moving Gratings.** **(A)**, Time courses of spatio-temporal activity patterns (top rows) and spatial averages (bottom traces) expressed as fractional change in fluorescence relative to blank condition (ΔF/F). Scale bar 1 mm. Here and in all other plots: *green* = responses to vertical grating, drifting rightwards in visual space; *blue* = horizontal grating, drifting downwards. **(B)**, (Top left corner) Singular values ranked in order of their contributions. Components of significant contribution to variance are colored (*gray* area depicts significance level). The contribution of each single SVD component to individual recording trials (n = 35) were computed, their correlations across trials are represented as a matrix. Spatial and temporal profiles of the first 6 significant SVD components (top row and left column) were clustered according to their correlation (*red* and *yellow*).

meaning to the individual SVD components or to any combination of a subset thereof. We reasoned that if given sets of components describe an underlying biological process, their relative contributions across trials should exhibit some degree of common variation. By measuring the correlation between the contributions of different SVD components across trials, we detected two clusters (3.5b, correlation matrix). The first cluster contained two SVD components (Fig. 1b, left column), each displaying non-oscillatory tonic activity profiles with highest singular values (red circles in 3.5b, top-left). The second cluster (3.5B, top-row) included the remaining 4 different SVD components with strongly oscillating activity. Their relative contributions were similar. However their weight was two orders of magnitude smaller than the first SVD components, suggesting weak contribution to the overall evoked activity.

Fig. 3.6 illustrates the spatio-temporal dynamics described by these two clusters of non-overlapping SVD components. Reconstruction of activity using the first two SVD components displayed the typical patchy structure of orientation maps characterized by repeating local domains of peak amplitudes (see contours in red panels and compare to

3.5). Note that the gratings evoked the expected orthogonal maps of activation. Overlaid on these maps, the oscillatory SVD components revealed cyclic waves of activity that propagated either medial-laterally (rightward in image frames) or in posterior-anterior direction (downwards) across the cortex depending on the gratings' drifting direction (contours in yellow panels). These propagating waves were brought about by oscillations with a period of ≈160 ms. The same patterns were observed in two further experiments. Thus, responses to each of the gratings' motion direction showed a clear shift over time indicating stimulus locked retinotopic propagation. Note that the distance that the wave traveled across the cortex in one cycle was nearly similar for both conditions. This is due to the fact that the imaged region matches the part of the cortical retinotopic map for which cat A18 is approximately isotropic, 5-10 mm below the representation of the area centralis. In the anterior part of the image, propagation was less pronounced, most likely due to smaller amplitudes of oscillatory activity resulting in decreased signal-to-noise in this region.

Next, to establish the dependency of the observed oscillations on cortical location, the data were fitted by a harmonic function with a constant period. Fig. 3.7A depicts the raw activity over time at two selected pixel regions (black and gray dots in 3.7b). We extracted the phase of the harmonic functions for each recorded pixel and plotted these values topographically (3.7b). In order to estimate the speed of the propagating waves, we then calculated the change in phase as a function of cortical distance (3.7c). A value of ∼34 mm/s was derived that matched the drifting speed of the gratings (34 °/s) in visual space. Consequently, the oscillatory SVD components reflected the shifts of the gratings' stripes by coherently propagating waves of activity with high spatial and temporal accuracy.

3.3.3.5 Comparison of Evoked Activity Levels

We next aimed to characterize the operating regime in which these spatio-temporal dynamics took place. We define an operating regime as the first, second and higher order characteristics of the distribution of activity levels. Fig. 3.8A depicts the activity levels averaged across the whole recorded area and stimulus presentation period, it thus illustrates differences in the first order statistics of the activity levels recorded under these two stimulation conditions. The overall average activity during presentation of different natural movies was equal to $0.57 \times 10^{-3}\Delta F/F$ ($0.46 \times 10^{-3}/0.68 \times 10^{-3}\Delta F/F$, $p = 0.05$). A value of $0.7 \times 10^{-3}\Delta F/F$ ($0.58 \times 10^{-3}/0.821 \times 10^{-3}\Delta F/F$) was observed with grating stimuli. Due to relatively large variability in the overall activity levels across experiments, we first compared grating and natural conditions within each experiment by computing their differences and later averaged these differences. In 9 out of 11 experiments grating stimuli elicited stronger responses (Wilcoxon signed-rank test, $p = 0.004$). The difference between average activity levels reached values as high as 58% of the activity evoked by natural movies; the mean increase was 26% (9%/41%, $p = 0.01$). As the underlying distribution of activity levels can be of any form, the average value constitutes a rather underdetermined description of the underlying distributions. Therefore to complement this analysis, we focused on the higher tail of the distribution of activity levels and analyzed the percentage of the recorded samples lying above a given common threshold value (Fig. 3.8A, right panel). A value of 17% (16%/18%, $p = 0.01$) was consistently observed when the threshold was set to be one standard deviation

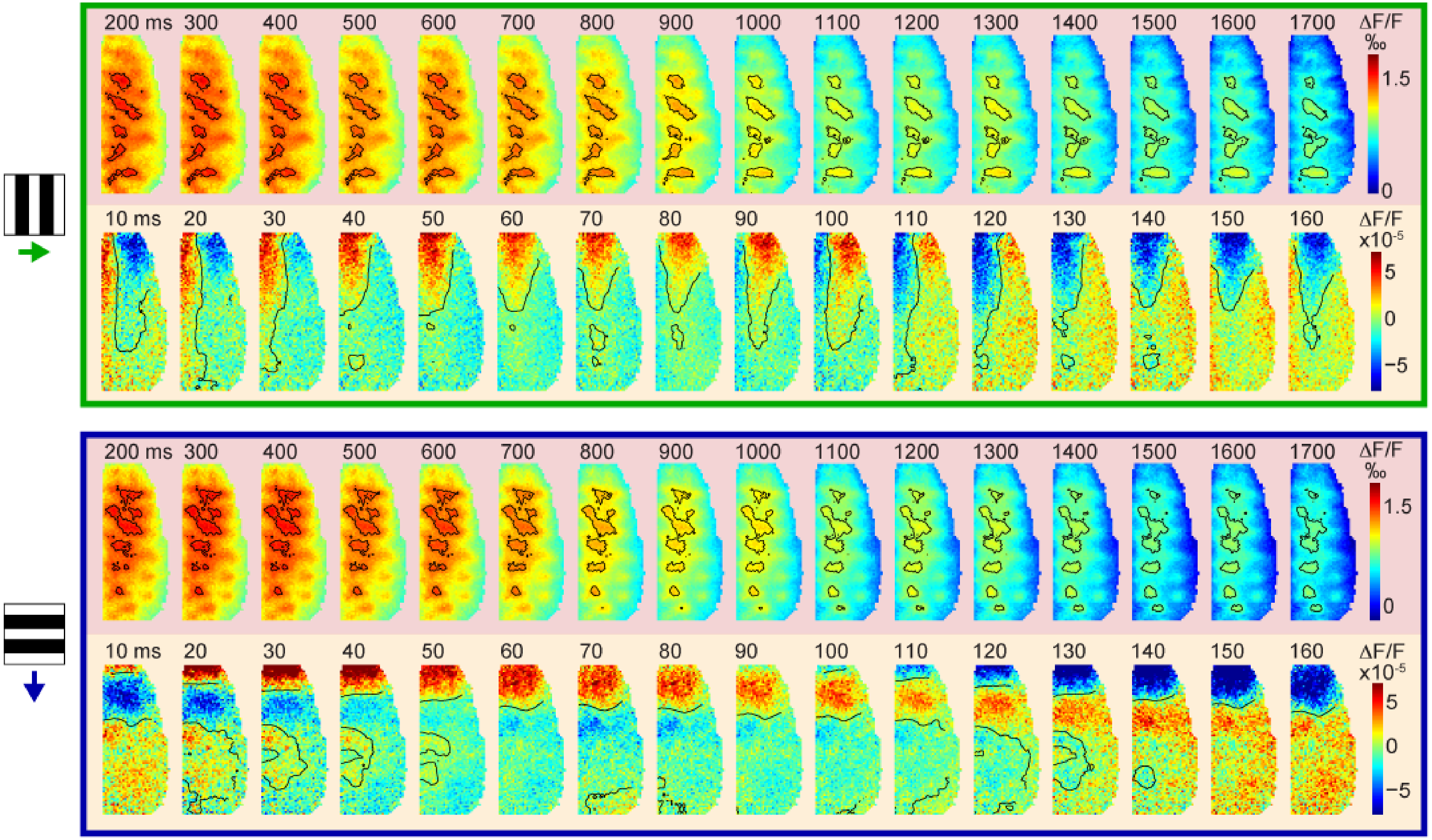

Figure 3.6: **Propagation of Activity across Stationary Orientation Maps.** Spatio-temporal activity dynamics represented by tonic and oscillatory SVD components (*red*, *yellow* panels, respectively). Icons at left sketch stimulus conditions. Time after stimulus onset indicated on top. Each frame represents average activity within 100 ms (*red* panels) and 10 ms (*yellow* panels). Contour lines are drawn around 90^{th} percentiles of activity for the tonic components and at zero crossings for the oscillatory components (multiple cycles of propagation were averaged in order to increase the signal-to-noise ratio). Colorbars depict activity levels, $\Delta F/F$. Note the difference of two orders between the two scales.

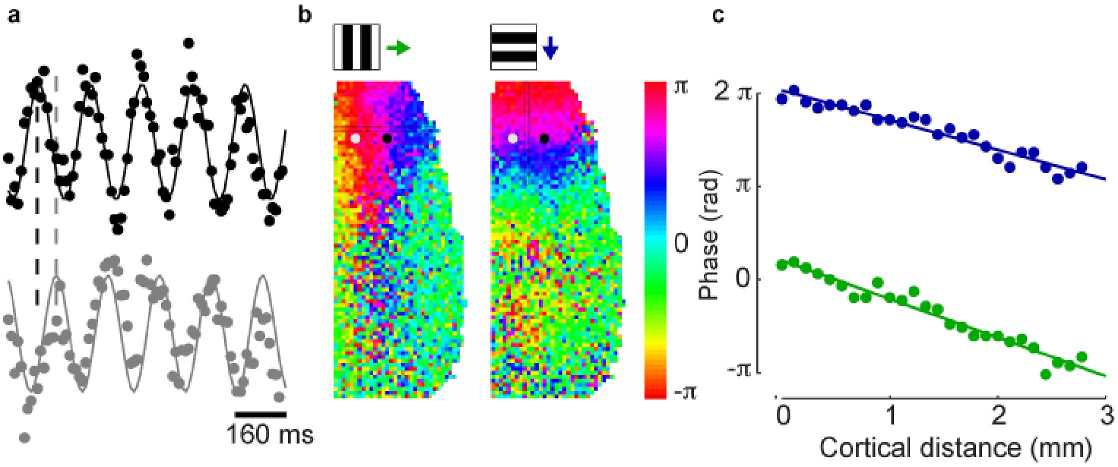

Figure 3.7: **Propagation Speed of Cortical Activity.** **(A)** Oscillatory activity around two pixels separated by 1 mm along the medial-lateral axis (*gray* and *black* dots in panel **B**). Oscillatory activity was best described by a harmonic function with a period of ≈160 ms (*black* and *gray* fitted curves). **(B)** The phase of each pixel is shown topographically for both grating conditions. **(C)**, Change of phase as a function of space within circumscribed regions (see vertical and horizontal rectangles in (B). The slope of the fitting line was ≈1.2 rad/mm.

above the average activity recorded during grating conditions (Fig. 3.8A, black dots). The median percentage of pixels lying above the same absolute threshold value was only 5% in the case of activity distribution evoked by natural movies (Fig. 3.8A, green dots). This suggests that the net excitation levels reached commonly during grating conditions were attained much more rarely during the processing of natural movies.

3.3.3.6 Specific vs. non-Specific Activity

As we were simultaneously recording neurons spanning the whole orientation space, we were able to dissociate the activity into two components, namely the specific (SC) and non-specific (nSC) activity. We divided the recorded area into two sets by choosing the most active pixels occupying the highest 5th percentile during presentation of horizontal and vertical gratings (Fig. 3.8B, pixels delineated with black/blue contours). nSC activity represents neuronal activity evoked by a grating that is orthogonal to the preference of the neurons (Sharon & Grinvald 2002). The incremental change in activity that is due to the change of orientation to the preferred one is termed the modulation depth (MD). We found SC and nSC activity to be equal to $1.06 \times 10^{-3}\Delta F/F$ ($1.21 \times 10^{-3}/0.91 \times 10^{-3}$, $p = 0.05$) and $0.88 \times 10^{-3}\Delta F/F$ ($1.02 \times 10^{-3}/0.75 \times 10^{-3}$, $p = 0.05$) respectively (Fig. 3.8B, middle panel blue and black bars). This corresponded to a MD of 20% (18%/24%, $p = 0.05$) of the nSC activity. The average activity during natural conditions across both sets of pixels was found to be equal to $0.74 \times 10^{-3}\Delta F/F$ ($0.64 \times 10^{-3}/0.88 \times 10^{-3}$, $p = 0.05$) and interestingly this value was smaller than nSC activity (Fig. 3.8B, middle panel green bar). In a pair-wise comparison, we found that on average the nSC activity was 19% (7%/30%) higher than the activity during natural conditions (Wilcoxon sign-rank test, $p = 0.018$).

As shown previously natural movies induced continual modifications in the balance between excitation and inhibition over the entire course of movie presentation; one inter-

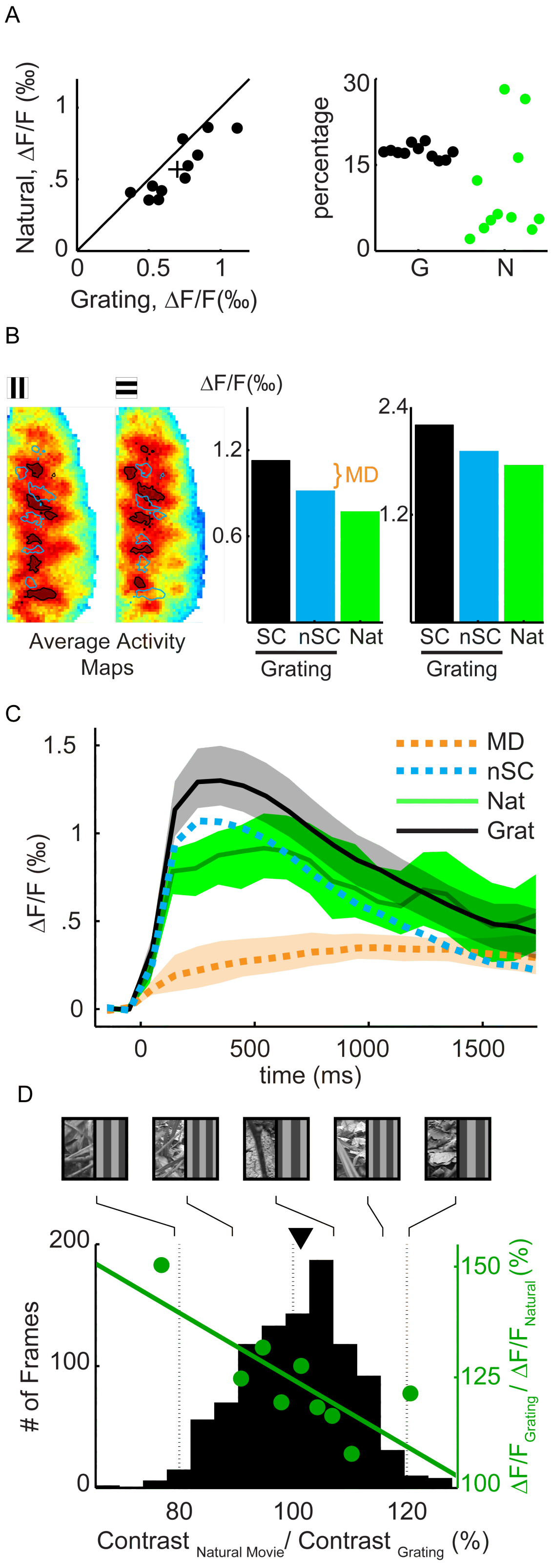

Figure 3.8: **Differences in the First-Order Characteristics of Response Levels.** **(A)** Left: Comparison of average activity evoked by natural movies and drifting gratings as a scatter plot for the complete data set ($n = 11$). Two natural movie (movie 1 and movie 2) and two grating conditions (horizontal and vertical) were averaged for each experiment beforehand. Plus sign represents average and SEM. Right: The percentage of samples that lies above a given common threshold for each experiment. The threshold was set to be 1 standard deviation above the mean activity computed during grating conditions (*black* dots). Same absolute threshold was used for natural conditions (*green* dots). **(B)** Single condition maps obtained for V and H gratings (averaged over the whole period of stimulus presentation). Pixels with orientation preference matching to the stimulus are delineated with *black* lines; their activity corresponds to the specific activity (SC). They occupy the highest 5th percentile of the activity distribution. Non-specific activity (nSC) is represented by *blue* lines which outlines the pixels with orthogonal orientation selectivity. Scale bar represents 1 mm. The middle bar plot depicts the SC (*black* bar), nSC (*blue* bar) activity averaged across all pixels and experiments. The difference between SC and nSC is termed Modulation Depth (MD). *Green* bar depicts the average activity within both sets of pixels during presentation of the natural movies. Rightmost plot shows the same results computed with peak instead of mean activity. **(C)** Temporal evolution of activity during presentation of grating and natural stimuli (*black* and *green* filled lines). 22 (Two grating and natural movie conditions per experiment) different individual time-courses were used. As various natural movies had different dynamics, average responses are necessarily less structured than individual time-courses. *Shaded* areas represent 99% bootstrap confidence intervals. *Dotted* lines represent the time course of MD (red traces) and nSC (*blue* traces) making up the response to grating stimuli. **(D)** Distribution of luminance contrast values within constricted regions (e.g. *white* frame in Fig. 3.2A) of movie frames which directly stimulates recorded cortical area. The values in the x-axis ranging between 65% and 128% represent the luminance contrast in percentage of grating stimulus contrast. The y-axis on the left represents the number of frame with a given contrast value. Example frames with increasing luminance contrast from left to right are shown on the top part. The triangle denotes the mean of the distribution which is slightly higher than 100% (Wilcoxon-Signed Rank, $p<0.05$). The distribution was divided into 9 intervals of equal number of frames, and for each frame the deviation from the activity evoked by gratings were computed (*green* dots). The median value of the deviation is depicted by the height of *green* dots. Best fitting line had an equation of ($r^2 = 0.61$, 0.94/0.11, $p < 0.05$) and crossed the x-axis at 132%.

esting question is to know whether the peak values that can be reached during natural conditions also differ from those reached during grating conditions. We therefore repeated the same analysis, this time based on the peak activation levels reached during the course of stimulus presentation (Fig. 3.8B, rightmost panel). The peak values averaged across experiments were approximately 2 times higher than the average amplitudes. Usage of peak activity levels resulted in a slightly reduced MD value of 14% (12%/16%, p = 0.05). This suggests that even during periods of very strong excitatory drive, the activity of neurons with orthogonal orientation preferences conforms to the functional cortical domains, albeit with a slightly reduced selectivity levels. We confirmed that nSC peak activity was higher than maximum values attained during natural conditions (Wilcoxon sign-rank test, $p = 0.024$) by an amount of 8% (2%/12%, $p = 0.05$). These results show that the processing of oriented bar stimuli is concomitant to a widespread vigorous excitatory drive; even sub-optimally driven neuronal populations were subject to a strong net excitation resulting in much higher activity levels encountered during stimulation with naturalistic movies.

3.3.3.7 Differences in the Time-Courses of Activities

In order to analyze the temporal evolution of the average activity, we computed the characteristic time-courses for both stimulation conditions (Fig. 3.8C, gray and green traces, shaded area represents 99.99% CI). These represent the activity averaged across different experiments and therefore illustrate systematic effects of the stimulation category rather than experiment specific influences (such as for example flow field locked specific temporal dynamics as shown in Fig. 3.3A). No global trends were observed in the time courses during stimulation with natural movies. On the other hand, despite their stationary nature as stimulation, the processing of drifting square-wave gratings was characterized by a strong adaptational decay in the global activity levels which led to a dramatic decrease of net excitation levels. To quantify this decay, we fitted a line to the global activity time courses and in order to exclude the transient part of the response, we only included samples starting from 300 ms after stimulus onset. In all experiments the grating stimuli gave rise to stronger decay values hence to more negative slopes and in no experiment a positive slope was observed. The slopes were computed after expressing the activity levels in percentage of peak activity recorded during grating conditions. For the grating condition, the slope was equal to -48%/s (-59/-35%/s, $p = 0.01$), meaning that after only one second of presentation time the amplitude of the net excitation declined to a level which is half of the peak activity levels. During stimulation with natural movies the mean adaptational decay was -24% (-44/-5%/s, $p = 0.01$), and the median slope was found to be only marginally different than zero (sign-test, $p = 0.065$). This suggests that the processing of natural movies, despite their complex spatio-temporal dynamics, is characterized by a rather stationary activity dynamics in the superficial layers of visual cortex. Given the fact that the natural movies we used were dynamic, these results are not surprising. Nevertheless, we would like to emphasize the fact that grating stimuli, which are the most common form of stimulation used in experiments, are processed in a regime which is strongly non-stationary with respect to continual change in the balance of excitation and inhibition.

We next analyzed the temporal evolution of the nSC activity and MD separately (Fig. 3.8C red and blue traces). We found that MD kept increasing during the first second of the presentation and reached a plateau relatively late and remained relatively constant until the stimulus offset (Fig. 3.8E). From 200 ms until 1000 ms the value of MD nearly doubled and reached values close to 0.35×10^{-3}. On the other hand the time course of the nSC activity (blue trace) was found to follow a very similar profile as the overall response time course (black trace) and toward the end of presentation it approached values as small as modulation depth. This similarity shows that the major characteristics of the temporal responses to grating stimuli is dominated by nSC activity whereas the activity responsible for the encoding of the stimulus properties such as orientation, has a much smaller and but also more stable temporal evolution.

3.3.3.8 Effect of Variability of Local Luminance Contrast

To what extent could local properties of natural images such as fluctuations in the luminance contrast (LC) values be responsible for the differences in the average activity levels observed? As stated above, the LC values within the local region of the natural movies directly stimulating the recorded cortical area exhibited some degrees of fluctuations (Fig. 3.8D, top row of frames). The distribution of local LC values is shown in Fig. 3.8D where 100% in the x-axis represents the LC of the grating stimuli which did not vary across time. LC values ranged between 65% and 128%. For each of the frames, we computed the evoked activity level associated with it and computed the ratio to the activity evoked by gratings at the same time of presentation. For example, to find the activity level evoked by a movie frame occurring at 500th ms, we measured the average activity at 590 ms and compared it to the activity evoked by grating stimuli recorded at exactly the same time. The median of the ratios were calculated after pooling all the data points which belonged to a given LC percentile interval (Fig. 3.8D, green dots) each containing the same number of frames. Not surprisingly the biggest difference between naturals and gratings was observed for the frames that had the lowest LC values corresponding to the interval of 65% and 88%; the median ratio of activity level reached values as high as 150 % (Fig. 3.8D, leftmost green dot). For the interval where the LC of the natural stimuli was approximately equal to the target contrast (percentile interval of 99% and 103%), the median ratio decreased down to only 120 % indicating that grating stimuli induced still stronger responses. The estimated parameters of the best fitting line using linear regression indicated that a boost of LC corresponding to 132% of the baseline LC level would be necessary in order to obtain the same activation levels between natural movies and gratings.

3.3.3.9 Deceleration/Acceleration Notch

We next focused on the transient part of the cortical response representing the arrival of the visual input to superficial layers of the cortical circuitry. The steep increase of the net excitation immediately following the stimulus onset was interrupted by a transient slowdown occurring over a short period of time (Fig. 3.9A, four examples). This has been referred to as the deceleration-acceleration (DA) notch in VSDI experiments and evidence suggests that it corresponds to the establishment of inhibition within the cortical circuits following the arrival of excitatory input (Sharon & Grinvald 2002). In all but two experiments we observed a notch occurring approximately 100 ms after stimulus onset. The exact time of DA notch varied slightly between experiments (Fig.

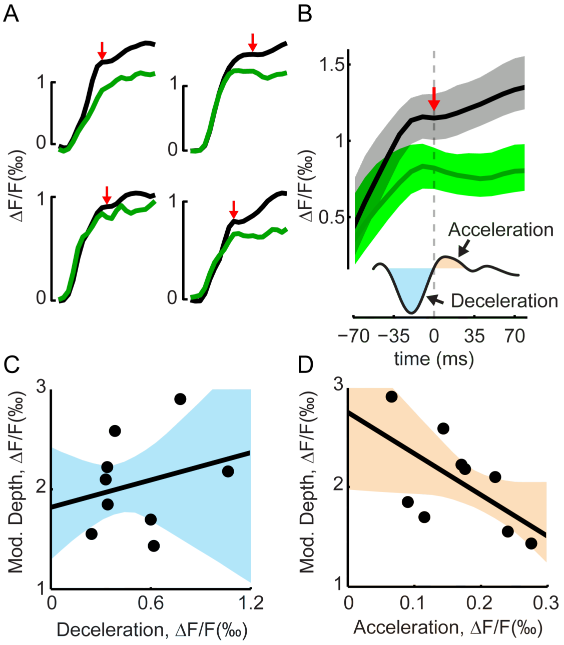

Figure 3.9: **Transient Part of the Response to Stimulus Onset.** **(A)** Four plots illustrate the deceleration-acceleration notch occurring during the transient part of the responses to gratings (*black* traces) and natural movies (*green* traces) onsets. The time runs from 0 to 200 ms. Deceleration was not strictly followed by acceleration during natural conditions. **(B)** Average across all experiments and 95% bootstrap confidence intervals. The time sample where the second derivative (shown below schematically) of the grating response crosses zero point was detected (zero point in x-axis and *red* arrows in (A) and (B)) and used to align in time different time courses prior to averaging. The area under the second derivative curve located in between the alignment point and the previous and next zero-crossings were used to quantify the strength of the deceleration and acceleration, respectively. **(C-D)** The relationship between the orientation selectivity quantified as the modulation depth and the strength of different DA notch components (deceleration in the left panel; acceleration in the right panel). 2 experiments out of 11 were excluded because the notch was not detected. Shaded areas represent the 95% confidence bands for the regression line. While the orientation selectivitiy was not found to be correlated with deceleration strength (r = 0.25, -0.41/0.68, p = 0.05), it was anti-correlated significantly with the acceleration component (r = -0.6, -0.96/-0.003, p = 0.05).

3.9A, red arrows). In order to align different experiments in time we detected the zero crossing of the second derivative of the dye signal (shown schematically in Fig. 3.9B) recorded during grating conditions (Fig. 3.9A, red arrows) and proceeded by averaging the time courses (Fig. 3.9B, black and green curves). The initial deceleration followed by a strong acceleration was evident in the average transient response in the case of stimulation with grating stimuli (Fig. 3.9B, black curve). A similar deceleration of activity occurred also in response to natural stimuli. However, the following acceleration component was much weaker or even absent for natural movies (green curve). As a consequence of the pronounced acceleration during grating stimulation, activation levels peaked at extremely high values; the activity between 150-450 ms was 48% higher than the overall mean, compared to only 23% in case of natural conditions. We thus conclude that the onset of grating stimulation caused an initial overshoot of activity within the first hundreds of milliseconds. On the other hand, the activity evoked by natural movies never or very rarely reached such high values. In order to check whether any bias in the statistics of natural images could be responsible of this outcome, we computed the luminance contrast within the area of the movies directly stimulating the recorded cortical area during the first 300 ms. The average luminance contrast was equal 100.29% (94.1%/105.2, p = 0.001) of the target luminance contrast. Therefore any bias in the contrast levels toward lower values that might be present in the natural movies within the first 300 ms cannot be held responsible for the discrepancies in the activity levels. These observations support the view that stimulation with gratings provides an excitatory input which is extremely powerful and affects the balance between excitation and inhibition in favor of strong excitation. In contrast, the system processes natural scenes in a more balanced regime leading to a marked absence of the late acceleration component and high activity levels. The extremely high activity levels hint into a qualitatively different state of the underlying cortical circuitries.

3.3.3.10 Strength of the Acceleration vs. Orientation Selectivity

We reasoned that the stronger the acceleration part, thus stronger the excitation opposing inhibition, weaker might be the selectivity of the visual cortex. To test this hypothesis we evaluated the selectivity of the visual cortex by computing the MD and quantified its relationship to the magnitude of different notch components across different experiments. The magnitudes of these were computed by integrating the area just neighboring the zero crossing of the second derivative of the raw dye signal (Fig. 3.9C-D, red and blue shaded areas). While the deceleration component was uncorrelated with the cortical selectivity (r = 0.25, -0.41/0.68, p = 0.05), the acceleration was strongly anti-correlated (r = -0.6, -0.96/-0.003, p = 0.05). In line with the earlier findings showing higher selectivity for neurons under natural conditions, our finding supports the view that the cortex under strong excitatory input looses its selectivity.

The MD represents the difference between activity levels of neurons that are strongly and weakly driven. Therefore it can in principle be used to quantify the dynamic range of neuronal populations. However a measure similar to MD is not easily derivable in the case of activity patterns evoked by natural movies because the natural stimuli is not clearly segregated into orthogonal orientation components making it difficult to track down differences in the spatial inhomogeneities of activity levels. To circumvent this problem, we computed the standard deviation of the activity values for each frame separately (20 frames representing each the average activity during 100 ms) and in order to obtain systematic differences between these two stimulation conditions we averaged the results across experiments (Fig. 3.10A). We found that the grating stimuli induced slightly stronger spatial inhomogeneities, however these differences were not found to be significantly different. These results show that the second-order statistics, contrary to first-order properties of activity levels are not qualitatively different between these two categories of stimulation. The dynamic range which is made use of during encoding of visual input is similar under natural and more simplified stimulation conditions.

3.3.3.11 Population Sparseness

We constructed histograms of the activity levels by pooling all recorded samples of all experiments (excluding samples recorded during prestimulus and the transient part) separately for the grating and natural conditions. Most remarkable difference between these histograms was the strong rightward shift seen for the activity distribution originating from the grating condition corresponding to the first order differences (Fig. 3.10B, black trace). Moreover differences in the extreme values also were observed; while lowest values were more abundant during natural conditions, the opposite was true for high-end values. We quantified the higher order characteristics of these distributions, namely the kurtosis. The distribution of activity levels during natural conditions was found to be slightly more leptokurtotic with a kurtosis of 3.32 (3.31/3.33, p = 0.05) whereas a value of 3.07 (3.07/3.08, p = 0.05) was obtained from the distribution of activity levels under grating conditions.

These histograms incorporate changes in the activity levels that occur over the stimulus presentation hence their shapes are to some extent dominated by average activity fluctuations. For example the decay in activity levels during stimulation with gratings

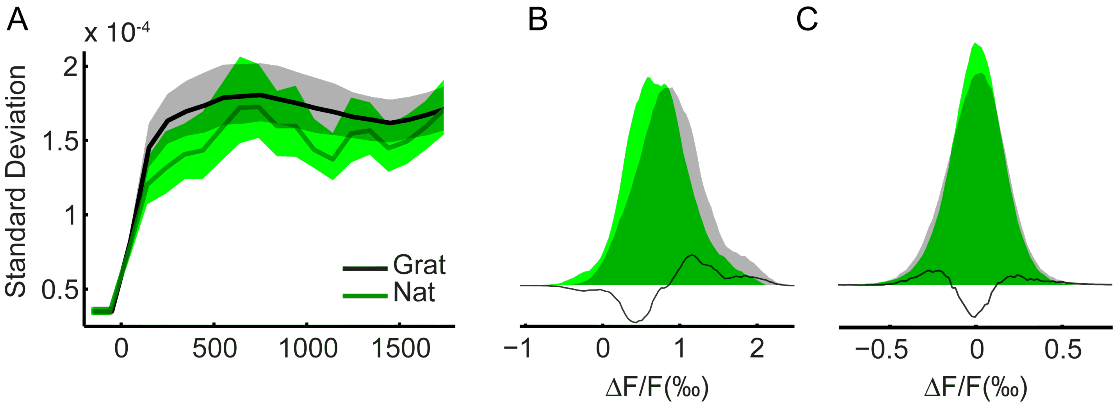

Figure 3.10: **Second and Higher Order Characteristics of Response Distributions.** **(A)** Temporal evolution of spatial inhomogeneity measured as the standard deviation of activity levels with a given frame. The values at each time point were averaged across different experiments after the standard deviation was computed. Grating stimuli induces slightly stronger spatial inhomogeneity, yet differences were not significant. **(B)** Histogram of activity levels of all recorded samples excluding the prestimulus and transient part of the responses for grating (*gray* shade) and natural (*green*). The data of different experiments are pooled together. *Black* curve depicts the difference between both histograms. **(C)** Same as in (B) but after removing from each frame its mean value. Both distributions deviate from a Gaussian distribution in terms of their kurtosis. However distribution during natural condition is more leptokurtotic with a kurtosis value of 3.91 in comparison to distribution obtained in response gratings which had a kurtosis of 3.35.

may have a non negligible contribution to the shape of the histogram depicted in Fig. 3.10B. In order to eliminate these confounding effects and to have a better overview on the spatial distribution characteristics of activity levels, we recomputed the histograms after removing from each frame its average activity level, thus centering all the frames around zero level (Fig. 3.10C). Both of these distributions deviated significantly from a Gaussian distribution in terms of their kurtosis values; a kurtosis of 3.35 (3.34/3.37, $p = 0.05$) was found for the spatial distribution of pixels under stimulation with gratings. For the natural condition, the distribution was more strongly leptokurtotic with a value of 3.91 (3.88/3.93, $p = 0.05$). These results shows that spatial distribution of activity levels evoked by grating and natural stimuli do have different higher order properties. Moreover high kurtosis values found under natural conditions is compatible with the view that neuronal populations have a sparse representation of the visual input.

3.3.4 Discussion

In the present section we address the question of whether and how sensory cortex processes simple, artificial stimuli versus complex, natural stimuli. Such a comparison between the stimuli most widely used in investigations of sensory cortex and the ecologically relevant stimuli is of major relevance for system identification. It is crucial to estimate in how far cortical functioning can be generalized across different stimulus categories. We investigated several basic measures of population activity: Peak amplitudes, average activity, stimulus motion locking and temporal characteristics, spatio-temporal separability and second-order and higher-order statistics. In order to simultaneously capture changes in cortical activity across several millimeters we used a state-of-the-art optical recording method that is sensitive to sub- and supra-threshold activity levels of a pool of neurons with a millisecond time resolution. We report large quantitative and qualitative differences in processing regime, indicating limited generalization across stimulus categories.

Natural stimulation evoked sparsely distributed population activity characterized by highly non-separable spatio-temporal dynamics that were locked to the intrinsic motion signals present in the movies. In contrast, responses to drifting gratings were characterized by a stereotypic rapid increase followed by a monotonic adaptational decay of activity. One of the major findings was the modest net activity level with balanced effects of inhibition and excitation in response to natural movies. This was in stark contrast to the vigorous excitation reached shortly after the transient notch during stimulation with gratings. An increase by almost half of the luminance contrast of natural movies was found to be necessary in order to equalize the activity levels between these different conditions. The cortical state during stimulation with gratings was dominated by a widespread feature non-specific activity and, importantly, even this non-specific activity was stronger than the average and peak activity levels encountered during the processing of natural movies.

In the interpretation of the high level of unspecific activity induced by grating stimuli and the lack of spatio-temporal separability of dynamics in response to natural movies we have to consider several methodological issues. First, due to highly reciprocal interconnectivity patterns within the superficial layers and the spatial extension of dendritic trees beyond functionally distinct domains (Gilbert & Wiesel 1989, Douglas & Martin 2004), the optically recorded signal, relating partly to subthreshold activity, could in principle suffer from averaging unspecific inputs. While, in the case of gratings, this would lead to an overestimation of the unspecific activity, it would also mask the non-separable dynamics when present during processing of natural movies. Indeed, it was shown that the majority of dendritic contacts within a radius of less than 500 microns do not exhibit an orientation bias (Malach et al. 1993, Bosking et al. 1997, Buzas et al. 2006). Thus, each orientation domain receives horizontal input from differently tuned neighboring neurons. These unspecific response components are evident in the orientation tuning of depolarizing responses, which is generally wider than for spike responses (Monier et al. 2003). This, however, applies more to natural movies that incorporate a rich spectrum of orientations and spatial frequencies (Simoncelli & Olshausen 2001). In the case of gratings the moderate bias of functional columns of similar preferred orientation to be connected (Rockland & Lund 1982, Gilbert & Wiesel 1989) and the properties of the reduced stimuli limit such an effect. Yet we observe a much higher level of unspecific activity in response to grating stimuli. This indicates that the high level of unspecific activity upon stimulation with oriented gratings, higher than the average activity induced by natural stimuli, is real and a sign of a qualitatively different processing regime.

Second, fibers of passage could in principle add to the measured activity at a given pixel, and therefore may lead to an erroneous overestimation of the amplitude of the unspecific response. However, it has been shown that the dye signal strongly correlates with the changes in membrane potential at the soma (Sterkin et al. 1998, Petersen et al. 2003b). Therefore it seems unlikely that activity at unspecific pixels is particularly infiltrated by specific signals originating from remote cross-orientation columns. Moreover our data reveal a large difference in the first order statistics of activity induced by gratings versus natural stimuli, yet the differences in second order statistics are smaller in comparison. These observations cannot be explained by contributions of fibers of passage. Instead voltage sensitive dye imaging constitutes a well-suited method for the detailed analysis of specific versus unspecific activation.

Along the same line, the fact that we observed highly non-separable spatio-temporal activity dynamics supports the view that optical imaging does no suffer from the aforementioned problems. The non-separability of activity during processing of natural movies resulted mainly from emerging spatial patterns that propagated across time and space and to a lesser extent by global changes in the activity levels such as adaptation. Notably, despite their simple statistical structure, the processing of gratings was also, albeit to a lesser extent, governed by non-separable spatio-temporal dynamics. The main source of non-separability in this case was the slight increase in feature selectivity observed across time accompanying the strong adaptational decay of unspecific activity. Additionally, we also observed propagating waves retinotopically representing the motion and the direction of gratings. In fact the existence of non-separable activity patterns has previously been demonstrated by in vitro measurements of neocortical patches (Xu et al. 2007) and during in vivo recordings of the visual cortex in response to simple stimuli (Prechtl et al. 1997, Benucci et al. 2007). Here we report that similar dynamics occur during the processing of natural movies. Therefore we conclude that the non-separable spatio-temporal dynamics is an important general quality of cortical responses to both simple and complex visual stimulation.

By means of relatively long recording duration of about 2 seconds, we were able to show that processing of continuously moving gratings takes place within a non-stationary cortical state. The monotonic decay in total activity during grating presentation relates to the well-known phenomenon of adaptation found in electrophysiological investigations (Maffei et al. 1973, Vautin & Berkley 1977, Movshon & Lennie 1979, Ohzawa et al. 1982, Kohn & Movshon 2003). Furthermore, also psychophysically the effects of long-term exposure to constant stimulation are well-studied (Gibson & Radner 1937, Blakemore & Campbell 1969, Pestilli et al. 2007). Moreover, adaptation is frequently used in functional magnetic resonance imaging experiments in order to reveal the selectivity of brain areas (Krekelberg et al. 2006). In all these experimental paradigms adaptation is commonly induced by constant stimulation lasting longer than 20 seconds, sometimes even many minutes. In our experiment, total activity was reduced by as much as 50% after only one second of stimulus presentation. This shows that within neuronal circuitries, major changes occur already very early under conditions of constant stimulation. This points to the fact that the cortical processing of the most commonly employed stimulation method takes place within a highly non-stationary cortical state. Interestingly, the decay in activity was not specific to neuronal populations receiving the preferred stimulus. Instead, as shown by the decay in the unspecific cortical activity, the entire recorded cortex, whether optimally or suboptimally activated, underwent the adaptational process.

As the major attributes of the activity in response to gratings were found to match the characteristics of the unspecific activity, it could be argued that the adaptational decay occurs independent of the specific content in the presence of “any” stimulus. For example, it could be a circuitry phenomenon in response to the sharp stimulus onset. However, for natural movies the adaptational decay was largely absent and on average only a small decrease in activity across time was observed. It is tempting to speculate, that one of the reasons might be, that in the case of natural stimuli the onset did not lead to a strong acceleration component, but stayed in a processing regime with balanced inhibition and excitation.

Alternatively the fluctuations in activity levels induced unceasingly by the dynamics inherent in natural movies may be held responsible for the virtual absence of the adaptational decay. The dye recordings revealed powerful modulations that were locked to the motion profiles of movies, and which accounted for a relatively large proportion of the variance present in the global activity time course. This implies that the observed motion-locking is a population phenomenon simultaneously affecting cortical regions of several square millimeters. Indeed, electrophysiological recordings have previously shown that under conditions in which stimulus velocity is not constant, local field potentials exhibit stimulus-locked modulations (Kayser et al. 2004, Schall et al. 2009, Mazzoni et al. 2008), which result in action potentials that are characterized by high-degrees of reliability across trials (Mainen & Sejnowski 1995). In contrast, responses to stimuli of constant velocity, such as drifting gratings, are characterized by a Poissonian relationship between the mean and the variance of spike counts (de Ruyter van Steveninck et al. 1997). Hence the low reliability of the spiking might be a result of the non-stationary temporal dynamics distinguished by the strong adaptational decay, and may therefore be a consequence of inappropriate stimulation used for identifying neuronal response characteristics. On the other hand, high reliability in responses may reflect an outcome of an adaptive phenomenon dealing with the ever-changing input patterns resulting from the animal's head and eye movements. The here observed activity levels of a large number of neurons locked to the motion cues is likely to underpin such a mechanism.

Shortly after onset of the grating stimulus, we observed a transient suppression of the rapidly rising activity, termed deceleration/acceleration notch in a previous report (Sharon & Grinvald 2002). The dye signal reflects net changes in potentials across membranes and therefore does not allow an isolated inspection of depolarizing and hyperpolarizing contributions. Thus the deceleration component, could in principle result either from withdraw of excitation or increase in inhibitory strength. However, withdraw of excitation is an unlikely scenario, given that the input was kept constant during both grating and natural movie conditions. It was previously argued that the notch results from the establishment of inhibition within the cortical circuitry following incoming excitatory inputs (Sharon & Grinvald 2002). It is presumably triggered by a strong decrease in membrane conductance due to shunting mechanisms (Borg-Graham et al. 1998, Hirsch et al. 1998, Sharon & Grinvald 2002). In fact, because activity within pixels coding for orthogonal orientation decelerated stronger than for preferred orientation, it has been interpreted as a signature of intra-cortical cross-inhibition (Bonds 1989, Ben-Yishai et al. 1995, Somers et al. 1995, McLaughlin et al. 2000, Shapley et al. 2003). In response to natural movies we observed a deceleration in activity with comparable properties, but the acceleration component was virtually absent. This resulted most probably from the effective inhibition leading to sharp termination of the increase in net-excitation. This provides evidence for a qualitatively different processing mode upon stimulation with gratings and natural movies. In the former case, excitation overcomes the rising inhibition in the network and activity reaches rather high levels. In the latter case, processing of the stimuli is performed in a regime of more balanced excitation and inhibition.

How does the increased effective inhibition detected in the dye signal relate to earlier results obtained at the single cell level? Upon stimulation with natural images, the strength of inhibitory subregions within RFs of LGN (Lesica et al. 2007) and V1 (David et al. 2004) neurons increased significantly compared to artificial stimuli. In line with these results, intracellular recordings have shown that for natural stimuli inhibition followed excitation within a much smaller time window than for gratings. Such rapidly counteracting inhibition leads to generation of fewer spikes and a more precise timing compared to gratings. Along the same lines, increasing the size of a natural movie such that it extends to the modulatory surround of the classical RF leads to highly non-linear changes in spiking behavior. The enlargement in stimulus size creates a net suppressive effect reducing overall population activity (Vinje & Gallant 2000; 2002), in line with our results obtained for movies that covered the entire visual hemifield. Moreover, in agreement with the modulatory suppression, natural stimuli produced increased sparseness (Vinje & Gallant 2000; 2002) and non-linear changes in feature contrast sensitivity (Felsen et al. 2005b). These observations underline the impact of suppressive mechanisms on cortical processing capacities and suggest its functional efficacy in a balanced regime of processing within neuronal circuits that are designed for natural input.

The hyper-excitatory state of the cortex upon stimulation with gratings may lead to erroneous overestimation of the bandwidth of visual feature tuning as the overwhelming synaptic activation in response to gratings that we have observed may cause wider tuning curves that occur during processing of oriented features in natural images. Supporting this conclusion we found the initial excitation wave to be detrimental with respect to cortical orientation selectivity as captured by modulation depth measure: During the first few hundreds of milliseconds where the net excitation peaks, the cortical spatial patterning due to orientation specific activity was weaker than predicted by the separability hypothesis. This resulted from a steady increase in the selectivity and in contrast to Sharon & Grinvald (2002), no brief transient increase in selectivity was observed. Instead, the strength of the acceleration was negatively correlated with orientation selectivity, indicating that inhibitory intra-cortical mechanisms, which sharpen orientation selectivity, were less effective with higher boosts of initial excitation. Therefore, we conclude that higher activity levels in response to gratings signifies that the operating point of the cortex was initially situated at a higher, ”artificial position”.

The qualitative differences in processing regime might be related to the differences in the statistical properties between these two classes of stimuli. Square wave gratings are characterized, as do natural images (Field 1987, Ruderman & Bialek 1994, van der Schaaf & van Hateren 1996, Torralba 2003, Betsch et al. 2004), by a $1/f^2$ power spectrum. Furthermore, the total energy carried by these stimuli was controlled to be the same. However there are considerable differences between these two stimulus categories. A grating is characterized by a single pure orientation and by infinitely long spatial auto-correlations, whereas spatial correlations in natural scenes are of finite length (Betsch et al. 2004). These regularities might result in a state where excitation surmounts inhibition, leading to higher levels of total activation. An alternative, fully compatible viewpoint interprets the different processing regime as an adaptational strategy to real-world stimuli. The reduced activity levels during processing of natural stimuli reduce energy consumption, an exceedingly important factor from an evolutionary perspective (Laughlin et al. 1998, Lennie 2003).

Our approach contrasted simple and artificial with complex and natural stimuli. Although these may be seen as extremes on a continuous scale, we observed qualitative differences. Gratings proved to be extremely powerful in driving cortical activity. For this very reason, cortical dynamics enters a regime, as signified by the deceleration/acceleration, with high total activity levels including a strong unspecific component. In contrast, under natural conditions dynamics are more complex, partly locked to stimulus dynamics and composed of many non-separable components. Whether the continuous adaptation, including the suboptimally activated regions, leads to a processing of grating stimuli comparable to the natural conditions with balanced excitation and inhibition remains to be investigated. From an evolutionary perspective, these observations might not be utterly surprising, given the necessity to limit the energy consumption of the brain to a level sustainable. Indeed, heightened attention to the processing characteristics in the regime of balanced excitation and inhibition is dearly needed (Felsen et al. 2005a). ;

3.4 Processing of Locally Presented Natural Movies and Contextual Interactions during Processing of Natural Movies

3.4.1 Introduction

In order to understand processing of natural stimuli it is essential to unravel neuronal interactions within cortical areas. An important characteristic of early visual cortex is long-range connections linking distant neurons to each other (Rockland & Lund 1982, Gilbert & Wiesel 1989). However their functions, especially under natural stimulation conditions are largely unknown. Here we investigated the dynamics of long-range cortical integration by using localized presentation of natural movies.

3.4.2 Experimental Procedures[‡]

In order to investigate long-range cortical interactions, natural movies were also presented locally through a single or a pair of Gaussian apertures. These were created by modulating the contrast of the movies as a function of space according to a two dimensional Gaussian function (FWHM $\approx 3°$-$4°$, size depend on distance of stimuli to area centralis in the individual experiments). The local movies were placed $\approx 3°$-$4°$ relative to the visual field projection of the area centralis.

Local conditions were indexed by two parameters: position (**A** or **B**) on the screen and movie number (1 or 2) specifying which full-field movies were to be masked. By displaying either one or simultaneously two local movies, the conditions **A**1, **B**1, **A**1**B**1, and **A**2, **B**2, **A**2**B**2 were created (Fig. 3.11). For conditions with a pair of patches the distance between the centers of the two Gaussian apertures was equal to 2.5×FWHM. In conditions **A**1**B**1 and **A**2**B**2 ,the stimuli played through both apertures were belonging to the same original movie (movie 1 or 2) and therefore these conditions are called coherent. Additionally we created incoherent conditions by presenting two different movies within both apertures. Local movies were corrected for mean luminance so that the average of the pixels within the Gaussian apertures was always equal to the brightness of the background they were embedded. However, we did not equalize the contrast within each apertures as it is not possible to do so without introducing strong artifacts, particularly in cases where the local portion of a movie frame contains a zone with homogeneous brightness values corresponding to a surface-like object in the movie.

For each experiment, a total of 8 local stimulation conditions were created (Fig. 3.11B). Using two positions (positions **A** and **B**) and two natural movies (movie 1 and 2), single (conditions **A**1, **B**1, **A**2 and **B**2), coherent (conditions **A**1**B**1 and **A**2**B**2) and incoherent (conditions **A**1**B**2 and **A**2**B**1) conditions were constructed. During stimulation with these localized stimuli, we observed evoked activity in 7 out of 11 experiments, which were included in the following analysis.

The data from 8 hemispheres (7 animals) were used for the results reported in the following (*local set*). The remaining data had to be discarded because we did not observe

---

[‡] This part complements the Experimental Procedures of the last Section. Here only the information specific to the localized stimulation protocol is provided.

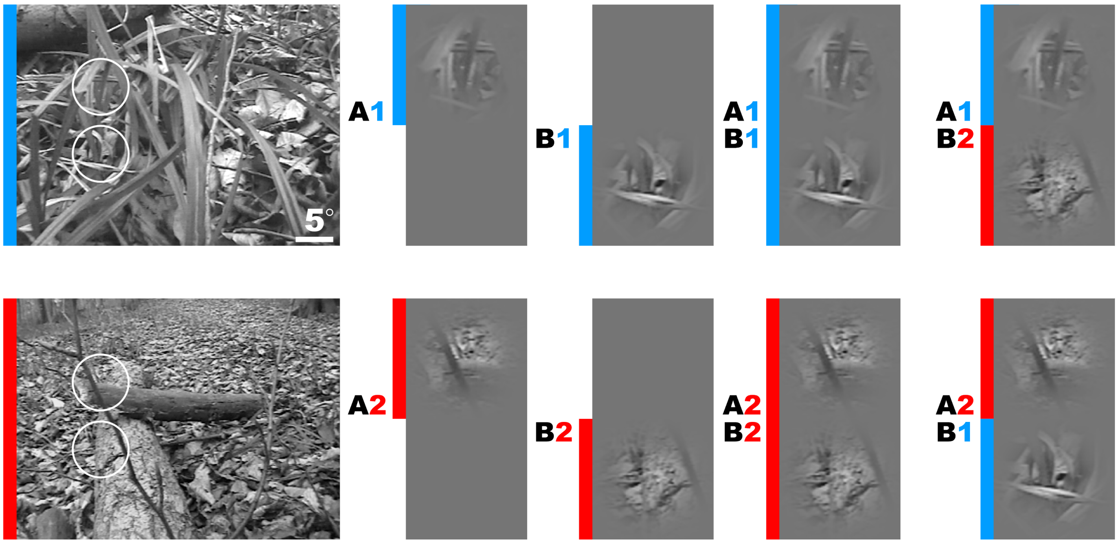

Figure 3.11: **Presentation of Natural Movies.** For each experiment, two base natural movies (marked *red/blue*) were used to create different stimulation conditions: In full-field conditions (left), movies maximally cover*ed* the visual hemi-field of the cat. In local conditions, movies were presented through either one or two Gaussian masks (3° or 4° FWHM, see Experimental Procedures) centered at positions "**A**" (top position) or "**B**" (bottom position), illustrated by *white* circles. According to the position label (**A** or **B**) and the index of the natural movies (movie 1 or 2), the following conditions were displayed: single (**A**1, **B**1, **A**2, **B**2), in which only one local patch was shown; coherent (**A**1**B**1, **A**2**B**2), in which two local patches belonged to the same full-field movie; and incoherent (**A**1**B**2 and **A**2**B**2), with local patches derived from different movies (see rightmost column).

any activity in response to localized presentation of the movies. Within this dataset local stimulation did not always activate the cortex strongly. We therefore had to distinguish between experiments with high activity levels from those where local stimulation did not yield a strong activity. Therefore the analysis were occasionally restricted to the experiments with high activity (*selected local set*, 3.4.2). The selection was done by applying a threshold on the peak average activity levels of the evoked data representing the average activity across presentation time and local conditions ($> 0.2 \times 10^{-3} \Delta$ F/F).

3.4.3 Results

3.4.3.1 Locally Presented Natural Movies Evoke Localized Dynamic Activity Patterns

Upon localized stimulation by natural movie patches, activity emerged from baseline level with variable delays among conditions (Fig. 3.12 shows responses to two different movies: movie one, rows 1-3; movie two, rows 4-6). The cortical responses were located along the cortical antero-posterior axis, reflecting the vertical positioning of the movie patches in the visual field. The localized stimuli gave rise to different spatial extents and peak values of cortical activation at different frames. Across the time averaged maps (Fig. 3.12, second column from right), spatially localized spots of activity with approximately centrally located peak values were clearly visible. These spot-like activation patterns observed during localized presentation allowed us to disambiguate the contributions of subcortical input from those of other sources. We were thus able to choose non-overlapping regions of interests (ROI) that were specifically driven by individual movie patches presented at the two different locations. The direct input to these ROIs was the same under coherent and single patch conditions, thus any differences in activity can be attributed to impact of long-range cortical interactions. By selecting the most active pixels corresponding to the highest percentiles during single patch conditions (**A**1, **B**1, **A**2 and **B**2), we chose two pairs of different ROIs (one pair for each movie) per experiment (Fig. 3.12, rightmost column). This yielded 28 ROIs for comparison (local set). Occasionally we focused on a subset of 19 comparisons corresponding to the ROIs with the strongest responses to the stimuli (selected local set). In the following, the latencies and activity levels within these selected ROI will be compared across conditions to identify contextual interactions.

3.4.3.2 Long-range Cortical Connections Reduce Response Time

We investigated the extent to which contextual stimulation reduced the time required for neuronal populations to reach a significant activation level. We defined this as the time it took for a recorded pixel to exceed two standard deviations of baseline in two consecutive frames, and visualized these values topographically using latency maps (see 3.13A, left and middle columns for examples). In the case of single conditions (left column), latency maps exhibited a localized circular spot embedded in a background that did not reach significant activation levels (black pixels). We also found a common tendency for the fastest reacting pixels to be centrally located, with an increase in latencies towards the periphery. To quantify the effect of contextual modulation on response times, changes in latency due to the additional movie patch were evaluated within a given ROI. To avoid problems with determining latencies of weak responses, the analysis was carried on the selected local set.

Within the highest percentile of activation (3.13A, red pixels in the right panel correspond to the 5th percentile) latencies exhibited a large variability, ranging between 100 and 700 ms (3.13B). The mean latency decreased in the presence of an additional distant movie patch, with 76% of cases displaying shorter latencies. The average latencies were about 340 ms and 268 ms for single and coherent conditions, respectively (t-test, $p < 0.05$). Furthermore, we found that the effect on latencies increased approximately linearly with decreasing activity percentiles (3.13C). For the least active pixels, corresponding to the interval of 75-80 %, the average change in latency was about 140 ms. This was twice the size of the effect found for pixels belonging to the highest percentile interval. Hence we conclude that the presence of a contextual natural stimulus reduced the response times of distant neurons to their direct subcortical input. Furthermore the least activated pixels benefited most from this effect.

3.4.3.3 Lateral Connections Exert Facilitatory Influences under Natural Conditions

We next evaluated the effect of contextual movies on activity levels. The time course of activity within the pixels corresponding to the highest 5th percentile is shown for the example experiment introduced previously (3.14A). The evolution of activity during the local presentation of two different movies (1st and 2nd rows) at two different cortical positions (left and right columns) is depicted. Note the similarity in temporal structure

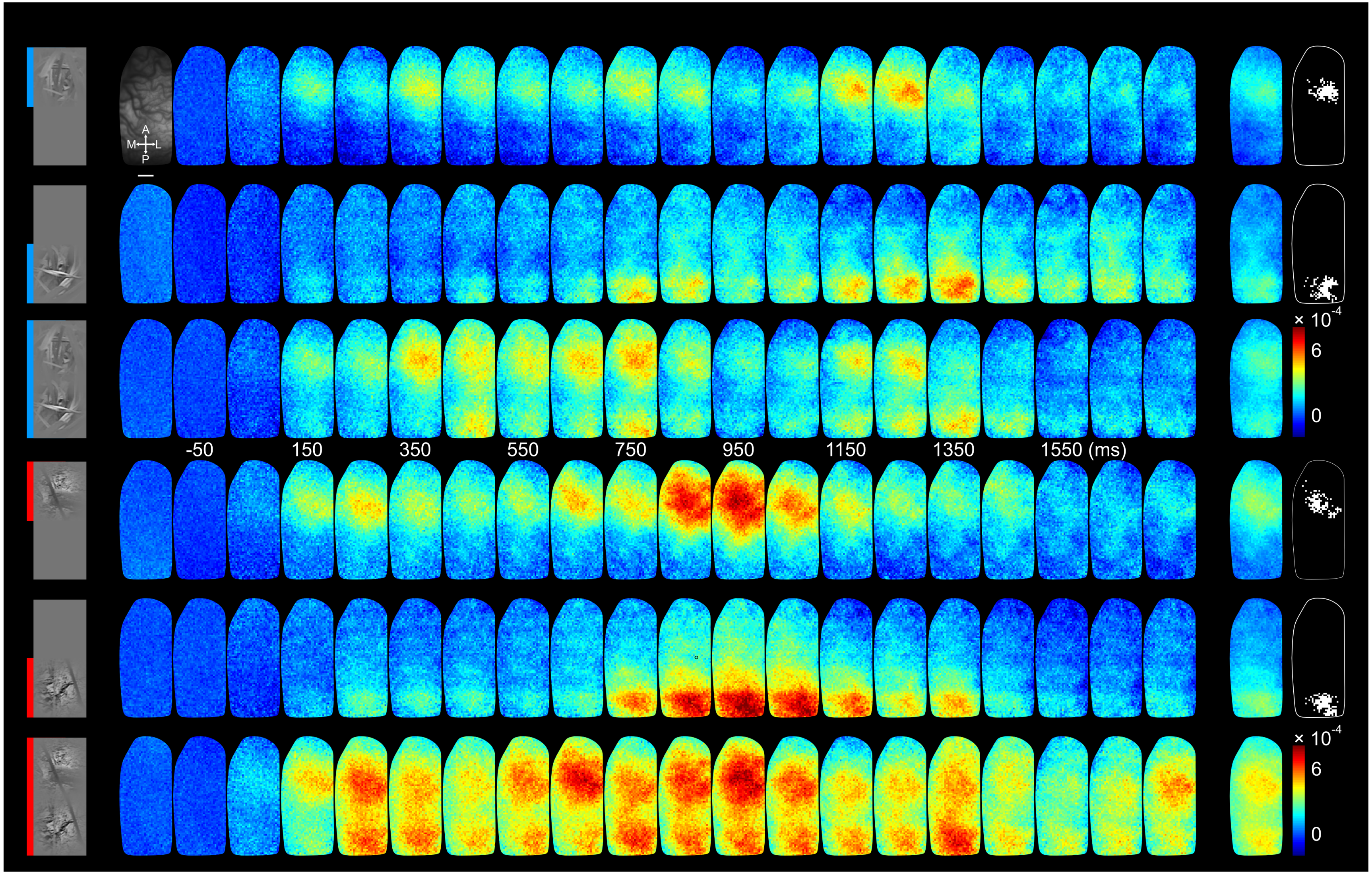

Figure 3.12: **Impact of Coherent Context on Cortical Activation Patterns.** Spatio-temporal activity patterns produced by locally presented natural movies. Leftmost column schematically shows the corresponding stimulus condition and the movie used (*blue* and *red*). Each single frame represents the average activity within 100 ms of non-overlapping segments of recorded data. Please note the differences in activity patterns between top and bottom triplets of rows where two different movies were used. The similarity between rows in a given triplet suggests that these differences are specifically caused by the individual dynamics and content of the movies. The column to the left of the colorbar shows the average activity maps computed across the stimulus presentation period. Rightmost column depicts pixels constituting highest 5th percentile of the entire activity distributions. Scale bar 1 mm.

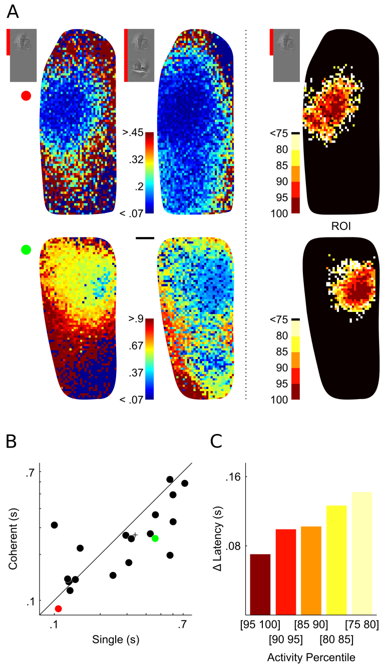

Figure 3.13: **Effects of Contextual Stimulation on the Latency of Activation.** **(A)** Latency maps derived from presentation of single and coherent movie patches. Color codes for the time point at which each pixel reached significant activation (non-responsive pixels in *black*). Left column: Latency maps during single movie conditions exhibit circular spots of latency values discernible from a background of non-activated pixels. Upper and lower rows show two different experiments. Middle column: Coherent conditions produce two spots in latency maps in which latencies are decreased compared to single movie presentations. Vertical bar marks 1 mm cortical distance. Colorbar denotes time after stimulus onset in seconds. Right column: To quantify the reduction in latencies as a function of activity levels, five sets of pixels were selected corresponding to five non-overlapping percentile intervals ranging from 95^{th} to 100^{th} until 75^{th} to 80^{th} based on average activity maps (not shown). **(B)** Average latencies computed across pixels within the highest percentile interval of activity (95^{th} to 100^{th}) are plotted for single and coherent conditions. Plus sign represents mean. Each experiment contributes 4 values (two regions of interests and two different movies). *Red* and *green* dots correspond to cases illustrated in **(A)**.**(C)** Reduction in latency as a function of activity levels.

within each row and the pronounced differences across different rows. This points to the fact that each movie has a specific impact on the recorded cortical region. The effect of contextual movies can be appreciated by comparing dashed and solid curves within each plot. The activity evoked by the same input was similar in the presence or absence of the second movie patch. However, higher activity levels for coherent conditions were consistently observed during the first half of stimulus presentation (black lines). Altogether this suggests that the direct input dominates the dynamics, while long-range interactions lead to facilitation upon presentation of contextual information.

To test whether this reasoning holds for the entire dataset, we took the median of differences across all 28 comparisons (solid black curve in 3.14B). We found that the facilitatory effect started to evolve briefly after stimulus onset, reached significance after 500 ms (Wilcoxon signed-rank test, $p < 0.05$), and remained effective for the next 400 ms. Across all experiments average activities over this first 900 ms of stimulus presentation time spanned a wide range, starting from no activity to values reaching $.4 \times 10^{-3}\Delta\mathrm{F/F}$ (3.14C). In 68% of comparisons, coherent stimulation yielded a higher level of activity compared to single conditions (3.14C). To test how this facilitation depended on the efficacy of the movies in driving cortical activity, we restricted the same analysis to the selected local set (19 comparisons, depicted by filled squares in 3.14C). In the remaining 9 cases, activity was very low, and did not exceed $.2 \times 10^{-3}\Delta\mathrm{F/F}$ within the pixels corresponding to the highest 5th percentile across all local conditions during the whole period of presentation. Within the selected local set, the activity during coherent stimulation was greater in 88% of the cases and the facilitatory effect reached significance as early as 200 ms ($p < 0.05$, Wilcoxon signed-rank test), with a peak at around 600 ms before vanishing at approximately 900 ms (3.14D, solid black curve). Our results so far have provided evidence that upon context, long-range interactions within early visual cortex are operational under natural stimulus conditions, and allow for enhanced activity levels and shorter onset latencies in the presence of contextual input.

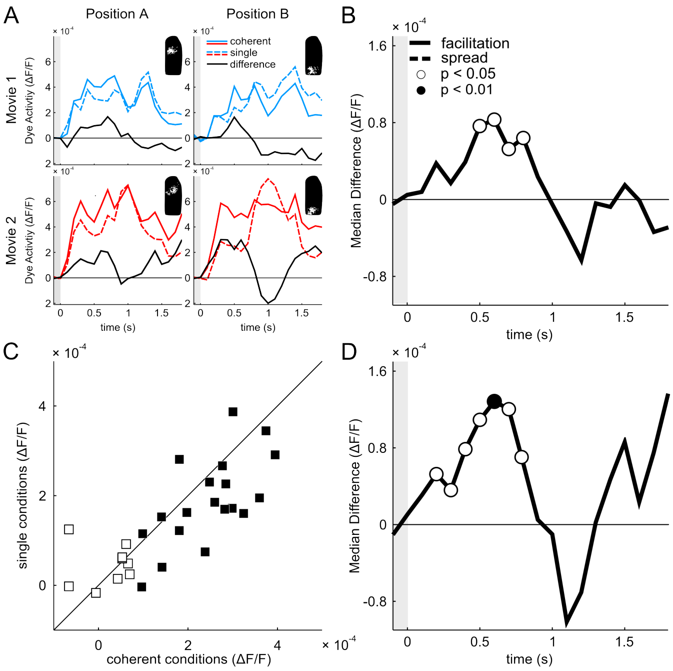

Figure 3.14: **Distant Natural Stimulus Exerts Facilitatory Effects on Local Activity Levels. (A)** Temporal evolution of activity during single (*dashed*) and coherent (*solid*) conditions for two natural movies presented locally (*blue*/*red* mark different movies; same example data as shown in Fig.3.12). *Black* line shows differences between coherent and single conditions. In all cases activity was initially higher in coherent conditions. The insets contain the binary image representing the pixels belonging to the highest 5^{th} percentile used for calculation of the traces. **(B)** Time course of the median difference between single and coherent conditions (*solid black* line) computed across all experiments (7 experiments, 28 comparisons). For each time sample, Wilcoxon signed-rank was used to test for deviations from zero (significant values marked by circles, see legend). Facilitation remained prominent over 900 ms. **(C)** Mean activity levels during the first 900 ms for single and coherent conditions shown for the entire dataset. In agreement with **(B)**, most of the data points lie below the diagonal indicating higher amounts of activity in response to coherent conditions than for single movie patches. *Filled* squares illustrate experiments in which the activities of the highest pixels during the whole presentation period and across coherent and single conditions were below $1.6 \times 10{-}4$. **(D)** Same plot as **B** summarizing experiments that revealed strong facilitation (*selected local set* see Section 3.4.2, *filled* squares in **C**).

3.4.3.4 Facilitatory Effects are Sensitive to the Spatio-Temporal Coherence of the Stimulus

In how far does the observed facilitation depend on the specific spatio-temporal properties inherent to the natural movies? As the movies were recorded from the viewpoint of freely moving cats, each natural movie had a distinctive flow field profile. Furthermore, motion was widely distributed across the visual field. During coherent conditions the same original natural movie was presented through both of the Gaussian apertures, and consequently the spatio-temporal characteristics of the natural movie were left intact. In contrast, in the case of incoherent conditions, movie patches stemmed from different natural movies, leading to an evident dissonance due to the elimination of naturally occurring correlations between apertures. The total input received by cortical neurons was identical across the sum of coherent and incoherent conditions (**A**1**B**1+**A**2**B**2=**A**1**B**2+**A**2**B**1), and therefore any difference encountered in activity levels found in this comparison can be attributed to contextual facilitatory effects arising from the spatio-temporal coherence of the stimulus.

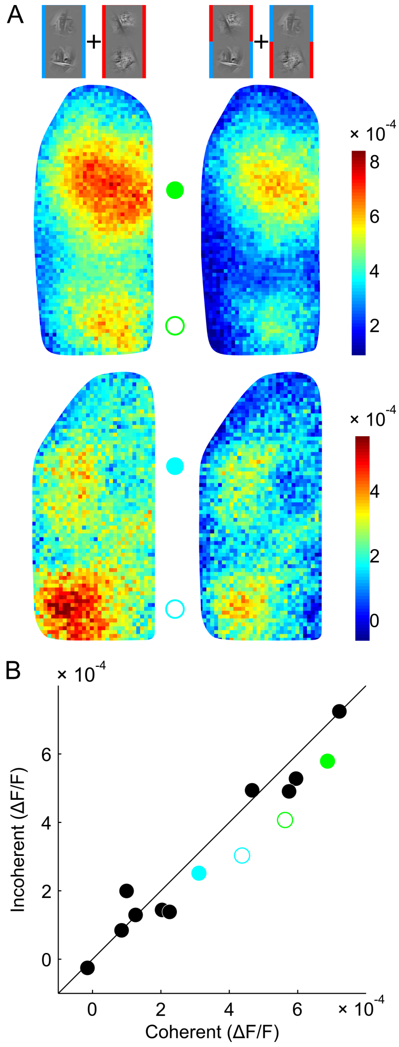

Figure 3.15: **Magnitude of Facilitation Depends on Stimulus Coherence.** **(A)** Cortical activation patterns evoked by coherent (left) and incoherent (right) movie patches shown for 2 different experiments (upper/lower row). Averages were computed across the first 900 ms in accord with the occurrence of long-range facilitation. Note that besides differences in peak values, the spatial profiles of activity in response to coherent conditions were enlarged; in particular cortical regions between the two activation spots showed elevated activity levels.**(B)** Summary across all experiments.

As facilitation was prominent during the early phase of stimulus presentation, we here focused on the first 900 ms. 3.15A shows average activity maps evoked by coherent (left) and incoherent (right) stimulation for two different experiments. As can be seen, the local movie patches evoked stronger and more spatially extended responses when the inherent spatio-temporal statistical properties of the natural movies were left intact. To quantify the differences in activity levels across all experiments we again compared average activity within those pixels that were most active (local). The location of these pixels corresponded to the peak position observed during coherent and incoherent conditions thus yielded two regions of interests to be compared per experiment. The average activity was in most cases higher for coherent conditions. This finding was consistent across the majority of experiments (3.15B), with 71% of experiments revealing higher activation levels for coherent conditions (Wilcoxon signed-rank test, $p < 0.05$). We thus conclude that the facilitatory effects are stronger when the provided context is coherent with respect to the spatio-temporal characteristics of the remote stimulation.

3.4.4 Discussion

Center-surround interactions are intrinsically tied to the integrative operationality of cortical function (for review Allman et al. (1985), Seriès et al. (2003), Albright & Stoner (2002)). In early visual areas, a fundamental property of neurons is their ability to sense regions of the visual space which is beyond their RF borders (Fitzpatrick 2000). The major characteristic of this ability is that isolated stimulation of surrounding regions is not sufficient *per se* to drive cortical neurons above firing thresholds. However strong modulatory effects are exerted on responses to a stimulus presented centrally, to the point of effectively changing tuning characteristics (Sillito et al. 1995). Previous experiments demonstrated a variety of facilitatory and/or inhibitory contextual effects but the final outcome crucially depends on the precise configuration of the parametrized stimulus used to stimulate center and surround regions. While the effect by surround was found to be mainly inhibitory (Hubel & Wiesel 1965, Maffei & Fiorentini 1976, Jones et al. 2001) and spatially asymmetrically organized (Walker et al. 1999), it has been shown that the precise nature of the effect depends on the contrast of contextual stimuli relative to the contrast threshold of the recorded neuron (Toth et al. 1996, Sengpiel et al. 1997, Kapadia et al. 1999). Although these neuronal phenomena are usually interpreted from the viewpoint of their benefits on naturally occurring visual tasks such as contour forming, figure-ground separation, object segmentation, or perceptual completion of occluded objects, it is not clear whether —and if so, how —these center-surround interactions studied with simple stimulus configurations extrapolate to complex dynamic conditions that mimic natural input.

By using ”keyhole-like” presentations of the original natural movies through either one or two distant Gaussian masks, we quantified the effect of surrounding stimulation on local activity levels. We provide evidence that contextual integrative mechanisms are indeed operative under natural stimulus conditions. The local movies evoked spatially confined activation spots with approximately circular shapes. Contextual stimulation led to higher activity levels within these regions compared to when the movies were shown in isolation. This facilitatory impact of the distant movie started within the first hundreds of milliseconds and lasted for nearly one second. The changes in activation levels were accompanied by reduced response latencies, most effective at cortical regions that were activated only slightly when single apertures were presented.

There are three idealized mechanisms, each based on different anatomical substrates, which could mediate the observed facilitatory long-range interactions. Overlapping feed-forward thalamo-cortical input could be a trivial explanation for increased cortical drive during stimulation with adjacent movie patches. However, there are a number of counter-arguments against this explanation. First, the pixels analyzed were separated by distances larger than the spread of direct thalamo-cortical projections (Humphrey et al. 1985, Bringuier et al. 1999). This is additionally supported by the fact that our latency measurements revealed spatially separated spots of equal response onsets for each movie patch, thus reflecting synchronous early thalamic input at two distinct cortical regions. Furthermore, the pixels included in our calculations were closely located around electrode penetration sites corresponding to RFs that were clearly offset in visual space. We therefore exclude a major direct feed-forward contribution as an account for our results.

Rather, the dense network of unmyelinated intra-laminar horizontal connections, linking distant neurons across several millimeters of cortical space, is a potential substrate for the observed long-range effects. It has been shown that neurons within the early visual cortex of cats receive diverse of subthreshold input combinations (Monier et al. 2003). As we observed subthreshold cortical activity dynamics far beyond the retinotopic representations of the individual movie patches, local activity may therefore be influenced by horizontal lateral input. In addition, feedback signals originating from higher visual areas with larger RF sizes than in the primary visual cortex may play another, additionally important role. In voltage-sensitive dye experiments, back-propagating waves of activity have been shown to be initiated in further downstream cortical areas as early as 100 ms after stimulus onset (Roland et al. 2006, Xu et al. 2007). Thus, these connections act fast (Lamme & Roelfsema 2000, Hupé et al. 2001) and are likely to mediate surround modulations spanning considerable distances in visual space, whereas lateral intra-laminar connectivity may account for modulations within shorter distances (Angelucci et al. 2002).

An important cornerstone of long-range facilitation is its dependence on the precise spatial configuration of the surrounding context (Nelson & Frost 1985). It has been shown that facilitatory effects decrease with decreasing congruency of the contextual stimuli with respect to the center stimulus (Polat et al. 1998, Kapadia et al. 1995). Using static stimuli, such coherence is generally controlled parametrically by changing the orientation difference between center and surround patches (Levitt & Lund 1997, Polat et al. 1998, Sillito et al. 1995, Chisum et al. 2003). As our stimuli were dynamic and complex, to control the coherency of the stimuli across two Gaussian apertures, we adapted a non-parametric method by exploiting the unique spatio-temporal characteristics of each original movie. When the same movie was presented through both apertures they were perceptually bound without effort, and appeared to belong to a single scene. On the other hand, the content within both apertures appeared to be immediately incompatible when two differing movies were used. There are a number of factors that determine coherence between patches taken from the same movie. First, stimulus motion was similar between the two distant apertures. This was due to the ego-motion of the recording cat, which induced large and equal motion fields across the visual scene captured by the camera. It has been earlier noted that such temporal phase relationships across distant regions are perceptually salient and enable object segmentation even in the absence of any spatial information (Lee & Blake 1999). Second, natural images tend to possess large spatial correlations because of the dominance of low spatial frequencies in their spectrum (Simoncelli & Olshausen 2001). Moreover, auto-correlations of orientations may cover large portions of visual field reaching up to 8 degrees (Kayser et al. 2003b).

We observed that the facilitatory contextual effects were slightly stronger when coherently moving movie patches were present in the context. Since the total input across coherent and incoherent stimulation received by the recorded cortical area was constant, these results cannot be solely explained by the local properties of movie patches. Rather, this facilitation necessarily reflects the outcome of an integrative phenomenon sensitive to the content of both local movie patches when presented simultaneously. Therefore, we suggest that the functional architecture of early visual cortical circuits may have empirically internalized the typical contextual relationships found in natural visual scenes.

4

An EEG Study: Motion Locking and Crossmodal Integration in Humans*

4.1 Context

In the previous Chapter (Section 3.3.3.2), we demonstrated that the dynamic natural stimulation has a major effect on the temporal profile of the cortical responses recorded in cats. By comparing the evoked optical activity waveform to the motion profile of dynamic natural movies we have shown that a good deal of the variations in brain activity is common to both of these signals. Moreover we provided evidence that the cortical machinery processing these dynamic input is able to integrate distant contextual information (Section 3.4.3.3) when the context is presented within the same sensory modality. In order to more fully understand the motion locking phenomenon (Kayser & König 2004) and to test its generalizability to humans, we designed a new set of experiments and benefited from the unequalized temporal resolution of EEG. Electroencephalogram is a well-established recording technique and offers recording of high frequency components (up to 100 Hz in our case).

To stimulate dynamically the visual system, we used a single Gabor wavelet. The choice of the stimuli was motivated by its well-parametrized nature. We modulated one of its parameters (for example its orientation) across time in such a way that the temporal trajectory was similar to the motion signals of our natural movies. Using this dynamic stimulation we investigated the motion-locking phenomenon at different frequency bands. Additionally, to test the effect of contextual information, we complemented the visual input with an extra-modal dynamic sensory signal. To this end, a frequency modulated auditory signal was used. The modulation of the frequency was either the same or different from the motion of the visual stimulation that was simultaneously presented. This

---

*This Chapter is based on the Master thesis of Sonja Schall (Schall 2008) which was supervised by Peter König and Selim Onat. A similar version of this Chapter is published (Schall et al. 2009).

resulted in a dynamic multimodal Gabor wavelet that was perceptually either congruent or incongruent across different modalities.

Our results showed that global brain activity, as recorded by EEG, locked to the dynamic stimulation at temporal frequencies way above those present in the dynamic stimulation. Moreover we demonstrated that extramodal auditory input has modulatory effects on the potentials recorded from electrodes reflecting presumably the activity of early visual areas.

These observations extend the results presented in the last Chapter: First of all we show that the motion locking phenomenon is not specific to the cat visual cortex but it can also be observed in humans. Moreover we show that the connectivity of the central nervous system subtends the integration of contextual information even when it is presented in a distributed way across different sensory modalities.

4.2 Introduction

Multisensory integration makes ecological sense when the incoming signals refer to the same external entity or, more generally, are due to the same underlying physical events. In general, there are three different ways in which signals must correspond for integration to take place: spatially, temporally, and/or semantically (Stein & Meredith 1993, Macaluso & Driver 2005, Calvert et al. 2001, Doehrmann & Naumer 2008).

These correspondences represent a default scheme of integration, and in special cases are influential enough to fool the system into binding streams that do not belong to the same underlying cause. In the well-known ventriloquist illusion, for instance, the concurrent temporal modulation of the puppet's mouth and the ventriloquist's speech leads to a perception of a talking puppet. In spite of the spatial and even semantic discrepancies here, temporal correspondence drives this audiovisual integration (Vroomen & de Gelder 2004, Bonath et al. 2007). A competition in temporal synchronicity between auditory and visual streams, on the other hand, may lead to a visual illusion. In the case of the sound-induced flash illusion (Shams et al. 2000; 2002, Mishra et al. 2007; 2008) an extra illusory flash is perceived when a single flash is interleaved between two beeps. More generally, at a constant visual stimulation rate, the number of perceived flashes increases with a higher auditory stimulation rate (visual illusion) and decreases when the rate of auditory events is slower (visual suppression) (e.g. Shipley 1964, Noesselt et al. 2008). The perceived temporal pattern of visual events is thus ”adjusted” to match the rate of auditory events. Taken together, these cases nicely illustrate that temporal aspects are essential to the perception of audiovisual events.

Research into the neural substrate of multisensory integration began in earnest after neurophysiological results showed characteristic multimodal response profiles at the level of single cells in superior colliculus in cat (Meredith & Stein 1983, Meredith et al. 1987) and macaque (Wallace et al. 1996). When stimuli were presented in close temporal proximity in two sensory modalities, the response of some collicular cells was found to reach or even exceed the sum of responses to each stimulus delivered in isolation. Imaging studies have since confirmed the importance of superior colliculus in human audiovisual integration (Calvert 2001, Miller & D'Esposito 2005). Several cortical areas have also been implicated in the processing of audiovisual stimuli including the insula (Bushara et al. 2001) and intraparietal sulcus (Calvert et al. 2001). The superior temporal sulcus, in particular, has been consistently shown to be an area of audiovisual convergence and integration (Beauchamp et al. (2004), see Calvert (2001) for a review of earlier work). Importantly, it has been found to be sensitive to the temporal synchrony of audiovisual information, especially audiovisual speech (Calvert et al. 2000, Miller & D'Esposito 2005, Macaluso et al. 2004), but also simpler stimuli (Noesselt et al. 2007, van Atteveldt et al. 2007). In addition, there is a growing body of evidence from ERP and fMRI studies demonstrating that areas traditionally conceived to be unisensory also play a role in the synthesis of audiovisual information (Giard & Peronnet (1999), Molholm et al. (2002), Kayser et al. (2007), Calvert et al. (1999), Miller & D'Esposito (2005), Noesselt et al. (2007), for general reviews on low-level integration see Foxe & Schroeder (2005), Kayser & Logothetis (2007), Driver & Noesselt (2008)). Another line of research follows the hypothesis that cortically, multisensory integration is instantiated by the relative timing of synchronized populations of cortical neurons. This has been supported by human EEG/MEG studies (see Senkowski et al. (2008) for a review).

To date, studies that explicitly investigate the temporal dependence of audiovisual binding, such as those mentioned above, have mainly manipulated the timing of stimulus onset or coincidence in one modality relative to the other. Typically, these experiments employ brief, simple stimuli that are either presented individually (Meredith et al. 1987, Bushara et al. 2001) or in streams (Calvert et al. 2001, Noesselt et al. 2007, Dhamala et al. 2007, Senkowski et al. 2007). However, such static audiovisual events are rarely found under natural conditions, where acoustic signals mostly emanate continuously from objects in motion, such as rustling leaves or moving cars. Real-world audiovisual events extend over time and are bound less by their simultaneity and more by their common temporal dynamics. For this reason, concentrating only on the isolated temporal coincidence of brief events neglects important temporal aspects inherent in natural events. To capture the full importance of temporal information, it is thus necessary to investigate audiovisual events extended in time.

Indeed, Kayser and König (Kayser & König 2004) have recently demonstrated that extended stimulation with natural visual stimuli leads to a continuous modulation of LFP power in cat visual cortex, and that this modulation reflects the dynamics of the presented movies. In addition, rhythmic auditory stimulation has been shown to entrain low frequency activity in auditory cortex of macaque (Lakatos et al. 2005). The entrainment of human cortical activity to stimulation at a constant frequency is also a well-known phenomenon in the auditory (Galambos et al. 1981, Bidet-Caulet et al. 2007) and visual modalities (e.g. Regan 1966, Ding et al. 2006), however it is not clear whether this also holds for stimuli with rich temporal power spectra. One possible role of stimulus locking in temporally-dependent multisensory integration may involve an increase in the efficacy of stimuli from a second modality when temporally aligned with neural activity that is entrained to the first modality (Lakatos et al. 2007, Kayser et al. 2008, Schroeder et al. 2008). As such, the extended changes in cortical activity seen in response to extended sensory stimulation are certainly worth closer examination in the context of audiovisual integration.

Here, we are interested in the general relevance of extended, time-varying information for the integration of visual and auditory streams. For this purpose, we paired well-defined visual and auditory signals that changed continuously and irregularly over time according to shared or differing dynamical patterns. Though relatively complex in temporal structure, and in this respect comparable to natural events including audiovisual speech, our stimuli have the advantage of being novel to the experimental subject and free of semantic reference. Using EEG, we employ two measures of stimulus locking to investigate audiovisual interaction. First, as a generalization of the evoked potential, we examine whether averaged EEG waveforms are locked to stimulus dynamics. Evoked potentials reveal the activity locked to stimulus onset, and here we extend this approach using cross-correlation. Second, following Kayser & König (2004) study, we investigate the locking of EEG power dynamics at frequencies within and beyond the range of our stimulus dynamics, and thus whether the observations made in cat transfer to man.

We specifically address three research questions. First, whether neural activity in the human cortex locks to the temporal structure of irregular dynamic stimuli in the visual and auditory modalities. Second and most importantly, we want to investigate whether multisensory integration is reflected in such a mechanism. Finally, we are interested in the time course of stimulus-locking, particularly in response to congruent and incongruent bimodal stimulation.

4.3 Experimental Procedures

4.3.1 Participants

EEG was recorded from 30 healthy university students, who were first screened for normal sight and hearing using standard tests (Landolt C chart and calibrated PC-based audiometer) and gave their informed written consent to participate in the study. Of these, 6 subjects were discarded prior to data analysis due to noisy data. The remaining 24 subjects (14 female, 23 right-handed) were aged between 19 and 33 years (mean: 23, standard deviation: 3). The experiment was conducted in accordance with the Declaration of Helsinki.

4.3.2 Experimental Design

In order to investigate multisensory integration processing, it is possible to contrast responses to unimodal and bimodal stimuli, or alternatively, to compare congruent and incongruent bimodal stimulation. To address both of these contrasts, we presented auditory and visual stimuli in four stimulation conditions: bimodal congruent, bimodal incongruent, unimodal visual, and unimodal auditory.

To engage their attention, participants were required to perform a congruency judgement task. In the bimodal case, subjects were asked to judge the temporal congruency of simultaneously presented auditory and visual stimulus components, thus ensuring that attention would be equally allocated to both modalities. To similarly engage subjects' attention during unimodal presentations, we adjusted the task to a sequential comparison of two consecutively presented unimodal stimuli, meaning that each such trial contained two stimuli. A sequential pair could consist of any combination of the two modalities (V-A, A-V, V-V or A-A). In subsequent EEG analysis, however, each stimulus of these sequentially presented pairs was treated as an independent unimodal stimulus in either

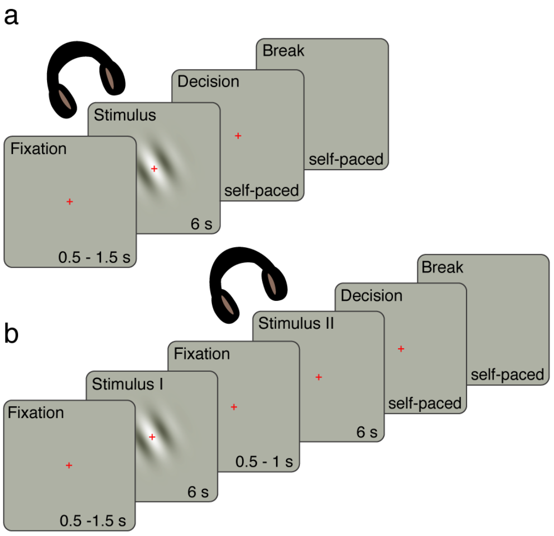

Figure 4.1: **Schematic diagram of both experimental paradigms.** For both paradigms trial onset was self-paced, and subjects indicated their readiness by button press. Each trial began after a random delay with the appearance of a red fixation cross, which subjects were asked to fixate until the end of the trial. **(a)** Bimodal trials consisted of simultaneously presented visual and auditory stimuli, and were either congruent (stimuli followed the same trajectory) or incongruent (different trajectories). **(b)** Single modality trials contained two sequentially presented stimuli, separated by a random interval (500 -1500 ms). A pair could entail stimuli from one or both modalities, again either congruent or incongruent. After stimulus presentation, subjects were required to decide whether the trial was congruent or incongruent, answering via button press.

the unimodal visual or auditory condition.

We chose a self-paced paradigm in which subjects triggered trial onset via button press (see Fig. 4.1). In doing so, we aimed to maximise participants' concentration and minimise eye-movements during stimulus presentation. Each trial began with the presentation of a red fixation cross. After a short, random delay, the stimulus appeared on the screen and/or was heard through headphones. In the case of bimodal stimuli, auditory and visual components were simultaneously presented, after which subjects completed the task by pressing one of two buttons (denoting "congruent" and "incongruent"). In the case of sequential trials, two stimuli were shown consecutively, separated by a short, random time interval (500-1500 ms), and subjects completed the task after the second stimulus. For purposes of behavioural analysis, task performance was evaluated on a trial basis.

4.3.3 Stimuli

Stimuli consisted of a rotating Gabor patch, a frequency-modulated tone, or both presented simultaneously, and lasted for 6 seconds. Gabor patches were grayscale, had a sinusoidal frequency of 0.33 cycles/°, subtended approximately 4.5° of visual angle (FWHM) and were centrally presented on a gray background identical to the mean luminance of the patch. Auditory stimuli had constant amplitude, and were ramped with a 10 ms half-cosine window to avoid clicks or other artifacts. Movies of visual stimuli were sampled at a frame rate of 25 Hz, and uncompressed visual and auditory stimuli were merged into an avi file for bimodal presentation using MEncoder (MPlayer version 1.Orc1, www.mplayerhq.hu).

Both unimodal stimulus components varied along only one feature dimension: orientation of Gabor patches spanned 180° centred around the horizontal axis, and the frequency of tones was modulated between 200 and 300 Hz. This ongoing feature modulation was determined by trajectories that changed smoothly but irregularly over time (see Fig. 4.2). Visual and auditory feature dimensions were chosen based on a behavioural pilot study (see Section 7.2 for more detail on how we have selected which feature to modulate during the main experiment) 32 trajectories were created in Fourier space by assembling a Gaussian power spectrum with a cut-off frequency of 1 Hz (resulting in a cut-off frequency of ≈4 Hz for stimulus speed) and random phase information. Each trajectory was constructed as part of a set of 4 mutually orthogonal trajectories, and in total 8 sets were created using an iterative optimisation procedure. The Fourier representation was then transformed to the time domain using the inverse fast Fourier transform. Stimuli were presented in congruent and incongruent conditions, where visual and auditory stimuli were modulated by identical or orthogonal trajectories (see below). This allowed us to precisely control for the degree of incongruence.

Stimuli were organised into four blocks, each consisting of 24 congruent audiovisual, 24 incongruent audiovisual, 24 visual, and 24 auditory stimuli (yielding 48 simultaneous and 24 sequential trials). Each block was created from two sets of orthogonal trajectories, with each trajectory occurring three times in each of the four conditions. Trajectories were balanced over conditions and presentation order of paired stimuli. Block order and stimulus order within blocks was randomised over subjects. Depending on time constraints and subject alertness, participants completed either 3 or 4 blocks (equivalent to 288 or 384 stimuli).

4.3.4 Presentation & Recording

Subjects were comfortably seated approximately 1 m from the presentation screen in a dimly-lit recording room. Stimuli were presented on a 21” monitor (Samsung SyncMaster 1100 DF) with a refresh rate of 100 Hz. The presentation software (Neurobehavioural Systems, version 10.3) was run on a Macintosh Pro (Apple Computers). Sounds were delivered via in-ear headphones (EAR-Tone 3A, Etymotic Research Inc.) at a comfortable volume chosen by the subject. During trials, subjects were asked to remain still and keep their gaze fixated on a centred, red fixation cross.

EEG was recorded from 28 sintered Ag/AgCl ring electrodes mounted in a cap (EasyCap GmbH, Herrsching-Breitbrunn, Germany) according to the standard 10/20 system, referenced to linked mastoids. Four additional electrodes were used to capture vertical and horizontal electro-oculograms (bipolar montages). The upper threshold for impedances was 5 kOhm. Two of the 30 standard electrodes (FT8 and FT7) were exchanged for diodes to capture visual and auditory stimulus onset. Signal amplification, filtering and digitisation were carried out by a 32-channel amplifier system (Synamps1, Neuroscan, Compudedics, TX, USA). The data were digitized at 500 Hz and online bandpass-filtered within 0.3 and 100 Hz to prevent aliasing. The digital signal was recorded by a PC-workstation (Intel Pentium 4, 2.41 GHz, 1GB RAM).

The diode signals revealed that the visual stimulus preceded the auditory by 25 ms —a small delay compared to the cut-off frequency of the trajectories —due to the stimulation software used.

4.3.5 Behavioural Analysis

Behavioural analysis served two purposes: first, to determine whether subjects were attentive during the experiment; second, to ensure that subjects could perceptually distinguish bimodal congruent from bimodal incongruent stimuli. Task performance was evaluated in terms of trials, and simultaneous and sequential stimulus presentations were dealt with separately.

To determine whether subjects performed above chance level, an exact Binomial test was performed for each subject. We furthermore used signal detection methods to quantify subjects' sensitivity and bias in discriminating between congruent and incongruent bimodal stimuli (see Wickens (2002) for more detail). The sensitivity measure *d'* was computed from hit (correct identification of congruent stimulus) and false alarm rates. Assuming that both rates come from standard normal distributions, *d'* estimates the distance between the two distributions in units of standard deviations —the further apart the two distributions, the easier the congruency can be detected, and the higher the *d'*. The response bias estimate ($\log \beta$) informs us whether the observer favours a 'congruent' or 'incongruent' response, with a negative value indicating a bias to congruence, and a positive value indicating an incongruence bias.

4.3.6 Pre-processing & Artefact Removal

All data analysis was carried out using MatLab (The MathWorks, Natick, MA), using the EEGLab toolbox (Delorme & Makeig 2004) and custom-made functions. Before artefact removal, all data were bandpass filtered between 0.5 and 100 Hz and epoched. Eye-movement artefacts were eliminated using Independent Component Analysis (Jung et al. 2000). Other muscle artefacts were detected by visual inspection, and whole trials were rejected. All trials were baseline corrected relative to 500 ms of pre-stimulus activity.

4.3.7 Stimulus Locking

Two approaches were used to quantify stimulus locking, both involving cross-correlations between EEG activity and the magnitude of change of the stimulus (see Fig. 4.2). All correlation coefficients were normalised so that the values lie between -1 and 1, with 0 indicating the absence of a correlation. Correlation coefficients are calculated for all temporal offsets between stimulus and response. Significant coefficients at positive time-lags indicate that the stimulus leads the cortical responses, that is the brain response is locked to the stimulus dynamics. Due to the finite autocorrelation of the stimulus, a locking might even occur at negative lags.

Our first analysis approach (spectrotemporal analysis) was concerned with the question of whether the time course of stimulus-induced power changes in the EEG correlates with changes in stimulus speed . This approach allowed us to investigate whether the power in any frequency band of the measured EEG signals locks to the dynamic stimulus.

A second analysis approach (waveform analysis) was used to examine whether the EEG waveform itself locks to the temporal profile of the stimulus. This approach was used to investigate whether there is a stable phase-relationship between evoked EEG response and stimulus. A significant correlation coefficient indicates that the phase of the EEG signal locks to the ongoing stimulus. In contrast to the spectrotemporal approach, this analysis only reveals locking of frequencies common to both stimulus and EEG. A detailed description of each approach is given below.

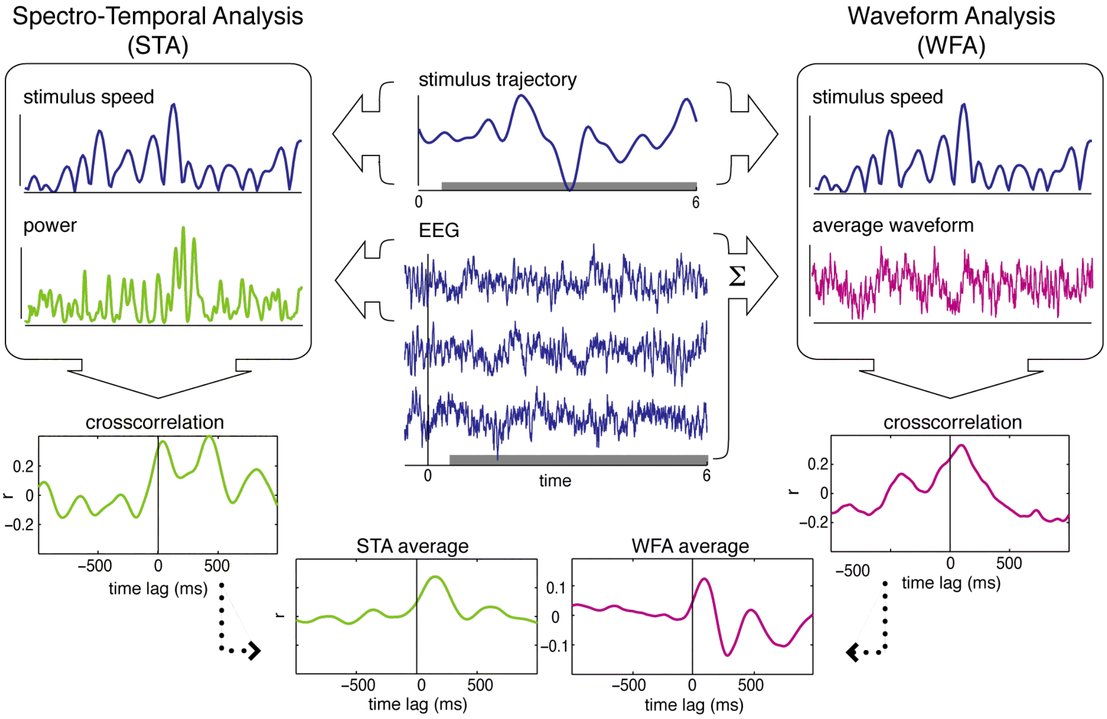

Figure 4.2: **Two Correlation Analysis Approaches.** Spectrotemporal and waveform correlation methods are illustrated on the left and right, respectively. Exemplary stimulus trajectory and single-trial EEG waveforms are shown in the centre, with a gray bar indicating the temporal range used in both analysis approaches. In the case of spectrotemporal analysis, each trial is transformed into a time-frequency representation (see text for details) and then correlated with the speed profile of the corresponding stimulus. The waveform analysis method involves creating an ERP from all trials corresponding to a single stimulus, and then correlating this waveform directly with the stimulus speed. Peaks at positive time lags indicate that the EEG power or waveform follows the temporal dynamics of the stimulus, in other words that the EEG has locked to the stimulus. In both approaches, correlograms are then averaged over stimuli. Grand averages are created by taking the median over subjects.

In the spectrotemporal analysis approach, spectral power changes in response to visual speed were determined on a trial-by-trial basis (Fig. 4.2, left panels). For each trial, we estimated the Short-time Fourier Transform using a window length of 80 ms shifted with an overlap of 78 ms. Each data segment was windowed (Hamming window) and zero-padded (to 256 samples ≈512 ms). Thus, we achieved a nominal temporal resolution of 2 ms and frequency resolution of approximately 2 Hz. The amplitude of the Short-time Fourier Transform was squared to obtain spectral power, and then z-transformed with respect to baseline power. Due to the power increase evoked by stimulus onset, we removed the first 500 ms of each trial which is uninformative for the present purpose and investigated only the stationary response. The power of individual frequencies was then either summed across our frequency range of interest (20-35 Hz) corresponding to previous results (Kayser & König 2004), or kept at its full resolution. Each frequency band was then separately cross-correlated with the speed of the stimulus, given by the absolute values of the trajectory's first derivative. The resulting cross-correlations were then averaged using Fisher's z-transform over all trials within each condition, yielding one cross-correlogram per subject, condition, frequency band and channel. Bimodal incongruent trials resulted in two cross-correlograms due to the difference in visual and auditory stimulation trajectories.

The waveform analysis approach evaluated whether stimulus speed is reflected in the EEG waveform by cross-correlating stimulus speed with stimulus-specific ERPs. An ERP was computed by averaging the EEG over those trials in which the same stimulus (i.e. identical trajectories) was shown for the modality in question (Fig. 4.2, right panels). Due to the limited number of trajectory repetitions throughout the experiment, ERPs were constructed from a maximum of 9 waveforms (please note that the number of trials is substantially larger). Stimulus-specific ERPs were then cross-correlated with the speed profile of the corresponding trajectory. The stimulus average for each subject, condition, and channel was obtained, again using Fisher's z-transform. As above, the grand average was calculated using the median.

Our first research question asked whether there is any evidence for stimulus-locking to visual and auditory stimulus dynamics, in other words how the cortex responds to the presence of a time-varying stimulus. To address this, for each modality we averaged over all conditions containing input to that modality, e.g. unimodal visual and both bimodal conditions were averaged to investigate visual stimulus-locking. This increased the amount of data used in our analysis, thus stabilising effects that were otherwise small and less consistent between subjects. In addition, we can thus estimate the modality-specific, automatic, bottom-up response irrespective of task factors or information present in other modalities. After evaluating the presence or absence of stimulus locking, the original conditions were used to explore the second question, namely whether locking is modulated by multisensory interaction. Finally, to address the third research question, we estimated changes in stimulus locking over the duration of a trial. This was done using a time-dependent cross-correlation function. Cross-correlations were computed within a 500 ms shifting time window (250 ms overlap), thus yielding 23 cross-correlograms per trial. Averaging over trials, conditions and subjects was performed as above.

4.3.8 Evaluating Statistical Significance

Significance of the resulting correlation coefficients was tested using bootstrap techniques. To test for significant peaks within a subject and for a given condition, we computed 1000 cross-correlations composed of averaged cross-correlations between non-matching EEG data and stimuli. From this distribution, the 95% and 99% confidence intervals were estimated. To test significance of the grand average, a distribution of 1000 medians was generated to determine confidence intervals.

To test for significant condition differences, permutation testing was used. For each subject, 1000 surrogate conditions were created by pooling and randomly redrawing trials from both conditions without replacement. A distribution of condition differences was obtained by subtracting the resulting 1000 grand averages (median) of both conditions.

4.4 Results

Here we first evaluate subjects' task performance in order to determine their alertness during the experiment and ability to distinguish audiovisual congruent from incongruent stimuli. Next we address the issue of whether stimulus locking can be found in the human brain. Finally, we compare congruent and incongruent bimodal conditions to investigate whether stimulus locking is involved in the crossmodal processing of temporal information.

4.4.1 Task Performance

All subjects performed well above chance in the task ($p < 0.01$, exact binomial test), with an average performance of 76 % (±8% standard deviation). This result indicates that subjects were attentive during the experiment and understood the task. Dividing data according to simultaneous audiovisual and sequential single modality trials, subjects' performance was better for the simultaneous (87 ±11%) than the sequential paradigm (65 ±7%). All subjects performed above chance in simultaneous trials (23 with $p < 0.01$, 1 with $p < 0.05$). In sequential trials the majority of subjects remained significantly above chance (14 with $p < 0.01$, 5 with $p < 0.05$), while 5 of the 24 were not able to significantly discriminate congruent from incongruent trials. The differences in performance are most likely related to the difficulty of the task —it is intuitively more difficult to compare two temporally extended patterns when they are serially presented than when they are simultaneously available.

In the case of bimodal trials, we further investigated subjects' ability to detect congruence and whether any bias was involved in their congruency judgement. This was necessary before proceeding with the contrast between EEG responses to bimodal congruent and incongruent stimuli. 22 out of 24 subjects had *d'* estimates greater than 1, with 19 of these exceeding a *d'* of 2. The subject average was 2.92, indicating that the difference between congruent and incongruent bimodal stimuli was clearly perceptually detectable. Interestingly, almost all subjects (21 out of 24) had a negative $\log \beta$ estimate revealing a bias to respond 'congruent' with a mean $\log \beta$ of -0.8 (±1 standard deviation). Maximum and minimum bias estimates were 1.65 and -2.6132, respectively.

As mentioned above, each stimulus of sequential pairs is treated as an independent stimulus in either the unimodal visual or unimodal auditory condition. Thus, we are not concerned with the behavioural results for those conditions. In all further analysis, all stimuli are used to calculate results, regardless of whether they appeared in correctly or incorrectly answered trials.

4.4.2 Stimulus Locking: Can We Find it in Human EEG?

We first had to establish whether EEG entrains to the dynamics of the presented stimuli. As mentioned in the Methods section, we quantify the locking for each individual input signal by correlating it with the measured EEG from any trial in which the modality of interest was stimulated with that input, thus averaging over conditions. Separate conditions are compared in the next section.

4.4.3 Visual Locking

The spectrotemporal analysis approach quantifies the amount of correspondence between the magnitude of stimulus change and the spectrotemporal power of the EEG within a given frequency range. To evaluate whether our data show stimulus locking, we first concentrate on the 20-35 Hz frequency band reported by Kayser & König (2004). Here we examine the effect at the population level, averaging first over all visual conditions (bimodal congruent and incongruent and unimodal visual) and then over subjects. As can be seen in Fig. 4.3A, we find stimulus locking to visual stimuli within this frequency range at occipital electrodes (O1, OZ, O2), peaking at a lag of 92 ms with a maximal correlation coefficient of 0.01.

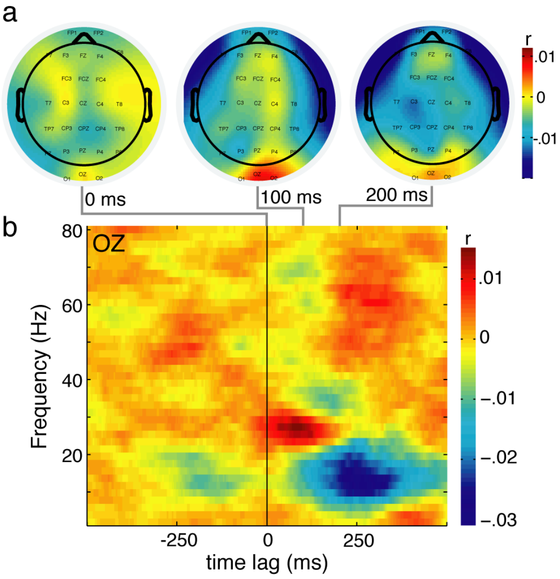

Figure 4.3: **Stimulus Locking of EEG Power to Visual Stimuli.** **(A)** Each plot shows the topographic distribution of the grand average of the spectrotemporal analysis correlation results (averaged over bimodal congruent, incongruent and unimodal visual conditions and over all subjects) within the 20-35 Hz band, at selected time lags beginning at 0 ms and ending at 200 ms lag. Dots represent locations of labelled electrodes, with locations below head centre drawn outside the cartoon head. Colour codes for the magnitude of the correlation coefficient, with warm colors indicating positive and cold colors negative correlation, and values are linearly interpolated between recording sites for visualisation purposes. **(B)** Grand average spectrotemporal analysis correlograms (as in a) for individual frequencies are shown for electrode OZ. Each row represents the results for a single frequency, with frequencies given on the y-axis and correlation lags on the x-axis. Colour codes for magnitude of correlation coefficient

Looking now at the entire range of frequencies that are available, again at the population level, we see at occipital electrodes that in addition to this positive correlation between visual stimulus dynamics and EEG power ($r = 0.015$ at OZ), there is also a negative correlation found in lower frequencies at a later correlation lag ($r = -0.03$ at OZ, see Fig. 4.3B). Although these effect sizes are small, they are clearly different to baseline, and significance testing using permutation tests reveals that these peaks are indeed highly significant ($p < 0.01$). Thus, we report two ranges of stimulus-locking: a positive peak centred at approximately 27 Hz occurring at 80 ms lag after visual stimulation; and a strong negative correlation between 8 and 20 Hz beginning at approximately 180 ms lag. The difference in direction of these correlations suggests that there are at least two frequency-specific stimulus-locking phenomena in response to visual stimulation.

These results are supported by an analysis of individual subjects. In 10 out of 24 subjects, a significant peak was found in the beta range for at least one occipital recording site. The clearest beta correlation was not found in the same occipital channel for all participants, with some displaying a strongly lateralized effect. The anti-correlation found in lower frequencies showed less variability across subjects, and EEG power of 20 out of 24 subjects was found to significantly anti-correlate with visual stimuli at an occipital electrode.

Visual stimulus speed is reflected not only in the power profile of single EEG trials but also in the EEG waveform itself, as can be seen from the waveform analysis results (Fig. 4.4A). The grand average cross-correlogram between stimulus speed and the stimulus-specific ERP waveform, averaged over all visual conditions, reveals peaks at various recording sites across the scalp. The strongest effects are found at occipital sites at 78 ms lag with a magnitude of correlation of 0.12. An extensive cluster of electrodes in the centro-parietal region shows a more complex pattern of locking, with a first positive peak at approximately -60 ms lag followed by a negative peak at around 180 ms and a second positive peak at approximately 390 ms lag. The strongest effect within this cluster is seen at CPZ ($r = 0.09$). Given that EEG at these sites follows the stimulus with differing delays, occipital and centro-parietal clusters are likely to reflect distinct underlying cortical sources.

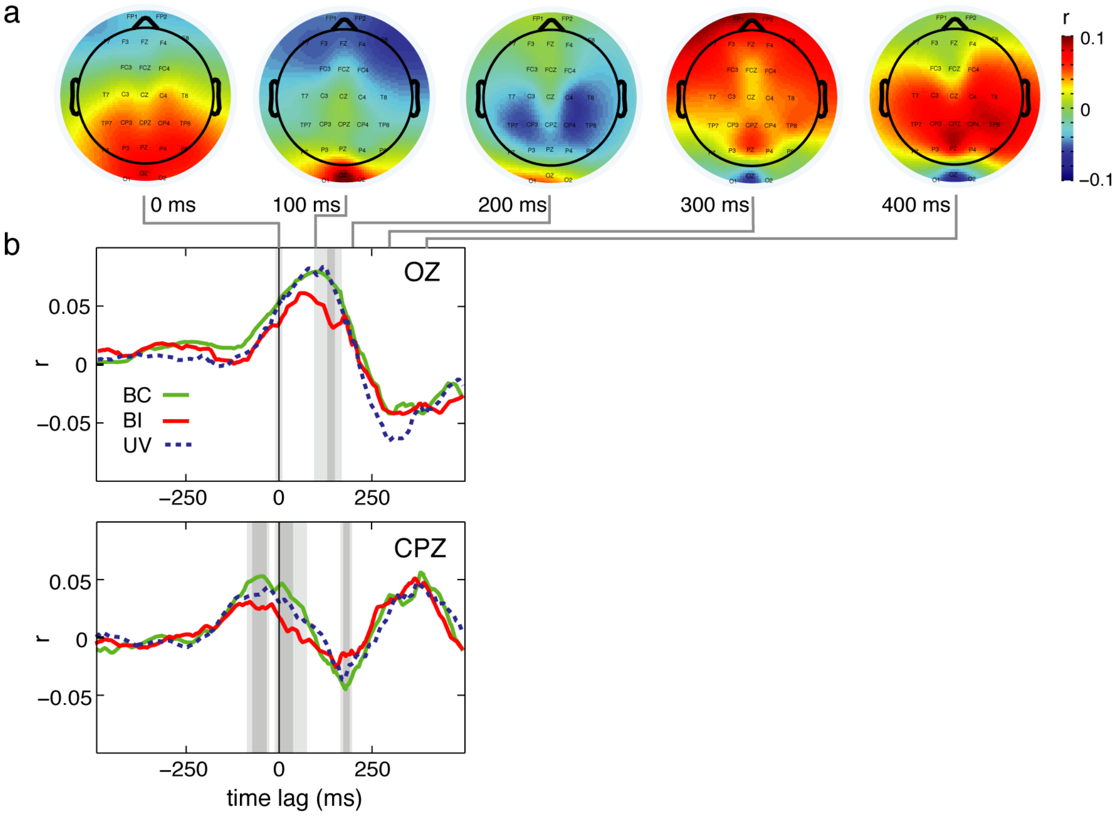

Figure 4.4: **EEG Waveforms Lock to Visual Input at Multiple Sites.** **(A)** Headplots show the topographic distribution of the grand average correlation coefficients between ERP waveform (averaged over bimodal congruent and incongruent and unimodal visual conditions) and visual stimulus speed at selected time lags between 0 and 400 ms. Correlation magnitude is color coded as in Fig. 4.2. **(B)** Condition comparison for electrodes OZ (upper panel) and CPZ (lower panel). Waveform analysis correlograms for bimodal congruent (BC, green line), bimodal incongruent (BI, red) and unimodal visual (UV, dashed blue) are shown with time lags on the x- and correlation coefficients on the y-axis. Time lags showing significant differences between bimodal incongruent and bimodal congruent conditions are highlighted in light gray for p<0.05 and darker gray for p<0.01

Contrasting the two analysis approaches, we see that the waveform analysis provides a stronger, consistent measure of stimulus locking. For almost all subjects (23 of 24), EEG recorded from OZ correlates with stimulus speed ($p < 0.01$ for 22 subjects, $p < 0.05$ for one subject). Significant peaks occur between zero and 300 ms lag (median r: 0.1372 for significant subjects, 0.1353 for all subjects) for the average of all visual conditions. Stimulus locking in the centro-parietal region is similarly stable, with 20 subjects showing a significant effect (19 with p <0.01, 1 with $p < 0.05$) at CPZ between 200 ms and 460 ms lag, (median r: 0.1229 for significant subjects, 0.1100 for all subjects). Thus, phase-locking, as captured by waveform analysis, is more reliable than the power-locking quantified by our spectrotemporal approach, making it more useful for investigating condition differences.

4.4.4 Auditory Locking

As a next step, we asked whether stimulus locking can also be found in the auditory system. Here, we didn't have an initial frequency band or site to guide our investigation. Representative results for both analysis approaches are depicted in Fig. 4.5 for electrode CZ (chosen because its auditory ERP showed the largest evoked potential of all channels shortly (100 ms) after stimulus onset) for the unimodal auditory condition. Although the spectrotemporal analysis results for this recording site suggest there may be stimulus-locking between 45 and 55 Hz, this correlation is not salient with respect to other lags and frequencies, and is furthermore not found at neighbouring channels nor is it consistent across auditory conditions. Waveform analysis cross-correlograms also show no salient peaks, with very low correlation coefficients over all time lags. As in the visual case, we also looked at the average over all auditory conditions, since we assumed that averaging over more trials might uncover small effects. However, the obtained results are similarly uninformative about auditory stimulus-locking. In addition, no consistent effects were seen on the individual subject level for either analysis approach.

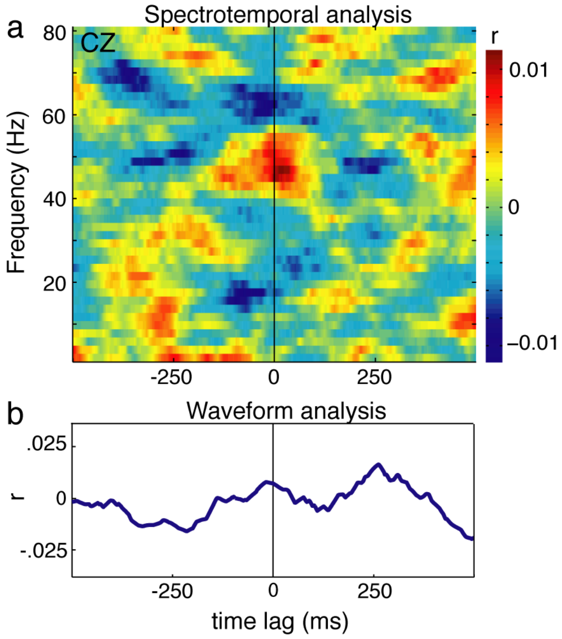

Figure 4.5: **No Stimulus Locking to Auditory Input.** Exemplary unimodal auditory results are depicted for electrode CZ. **(A)** Spectrotemporal analysis cross-correlograms are shown in rows for individual frequency bands (y-axis). Colours represent correlation coefficients. **(B)** The waveform analysis cross-correlation is represented with time lags on the x- and correlation coefficients on the y-axis

4.4.5 Is there Evidence for Multisensory Interactions?

To address the question of whether stimulus locking is subject to multisensory interactions, we contrasted different stimulation conditions. As we did not find any evidence for auditory stimulus locking, we concentrate in the following on visual stimulus locking and how it is modulated by the presence of matching or mismatching auditory input.

Despite the absence of auditory stimulus locking, the waveform analysis results revealed that incongruent auditory information modulates visual stimulus locking. This crossmodal effect is indicated by diminished peak correlation coefficients in bimodal trials containing incongruent auditory input, compared to bimodal congruent trials. Occipitally, the maximum correlation of the bimodal incongruent grand average ($r = 0.0612$) is almost 25% smaller than the peak of the bimodal congruent condition ($r = 0.0795$). Time lags between 98 ms and 168 ms have statistically less locking in incongruent than congruent bimodal trials ($p < 0.05$). The unimodal visual condition, in contrast, shows the same correlation as the bimodal congruent condition ($p > 0.05$). For centro-parietal EEG, the biggest condition differences are found for the first of the two peaks mentioned above. For time bins between -90 and -30, and -10 and 72 ms of lag, bimodal incongruent coefficients are significantly smaller than for the bimodal congruent or unimodal visual conditions ($p < 0.05$, see Fig. 4.4B). Likewise, the magnitude of the subsequent anti-correlation is lower for bimodal incongruent than both other conditions for lags between 168 and 198 ms (p <0.05). It seems then that incoherent auditory input suppresses visual stimulus locking during multimodal stimulation.

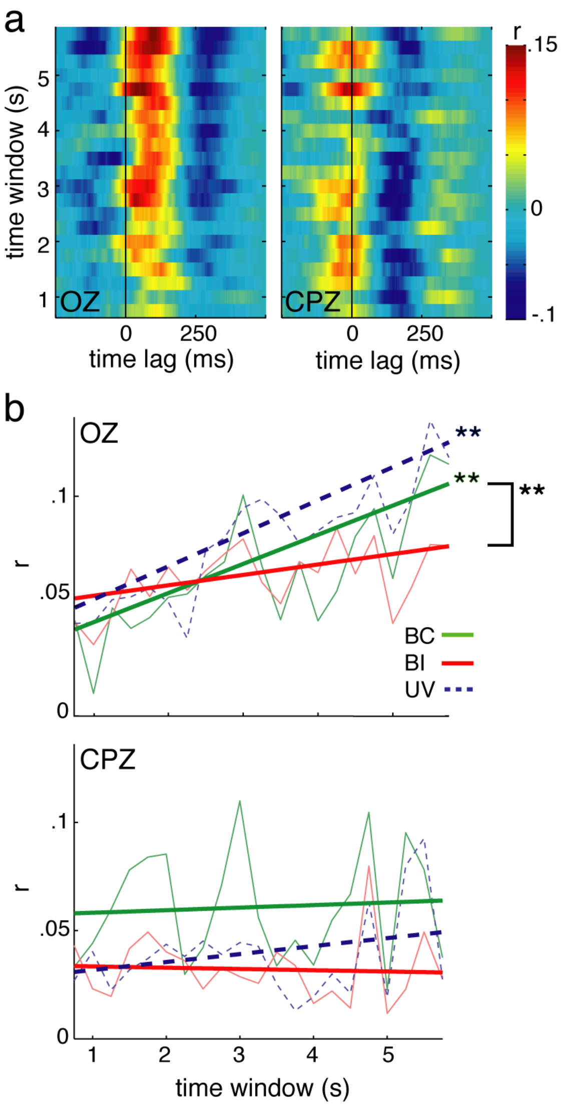

Figure 4.6: **Temporal Progression of Stimulus Locking.** **(A)** Each row shows one grand average waveform analysis cross-correlogram (waveforms calculated from average of bimodal congruent and incongruent and unimodal visual conditions), which was computed within a 500 ms shifting time window with 250 ms overlap. From bottom to top, the window is shifted from stimulus onset to the end of stimulus presentation. Correlation coefficients are color-coded. **(B)** The maximum waveform analysis correlation coefficients (y-axis) are plotted as a function of time for OZ (left panel) and CPZ (right panel). Peaks were determined within a region of interest (0-250 ms lag) for each time window from the grand average of each condition. The resulting data points are shown in light-colored lines for bimodal congruent (BC, green), bimodal incongruent (BI, red) and unimodal visual (UV, dashed blue). The according linear fits for each condition are shown in darker thick lines in the corresponding color scheme. Stars at line ends indicate significant linear trends (p<0.05) within a condition, and starred brackets indicate significantly different slopes between two conditions

Applying the spectrotemporal analysis method, no condition differences were found for the stimulus-locking effects mentioned above. However, the unimodal visual, bimodal congruent and bimodal incongruent condition all showed significant peaks in the frequency ranges and time-lags reported for the average condition reported above: peak cross-correlation values for the postive peak within the limits of 0-150ms lag and 20-30 Hz were 0.018, 0.020, and 0.17, respectively; for the negative peak they were -0.035, -0.034, and -0.035 within lag limits 150-500 ms and 8-20 Hz. Thus, spectrotemporal stimulus locking to visual stimulus dynamics seems to be a bottom-up driven cortical response independent of input to other modalities.

4.4.6 Do the Effects Change over Time?

To examine the progression of stimulus locking over the duration of stimulus presentation, we evaluated cross-correlations as a function of stimulus time. To this end, correlation coefficients were determined within a shifting time window (see methods in Section 4.3). Before investigating unimodal visual, bimodal congruent and bimodal incongruent conditions separately, we again examined their condition average.

Fig. 4.6A shows the waveform analysis results for selected occipital and central electrodes. At occipital recording sites, stimulus locking increases with progressing stimulus time —immediately after stimulus onset, correlation coefficients are very small and only start to increase after approximately 2 seconds of visual stimulation. In contrast, the magnitude of stimulus locking at central sites is not dependent on time. Locking is roughly constant, with no apparent systematic change. Overall, we see different time courses of visual stimulus locking at occipital and centro-parietal electrodes.

To quantify the temporal progression and obtain a better comparison across conditions, we extracted the maximum correlation coefficient for each time window. These time series were then fitted using linear regression. Results are shown for bimodal congruent and incongruent and unimodal visual conditions at selected occipital and central electrodes (Fig. 4.6B). At the occipital site, there is a significant linear increase for the bimodal congruent (regression coefficient = 0.0027, $r^2 = .5$; $p < 0.01$) and unimodal visual condition (regression coefficient = 0.0034, $r^2 = .75$; $p < 0.01$). In contrast, locking in the bimodal incongruent condition does not follow any linear trend (regression coefficient = 0.0007, $r^2 = .12$; $p > 0.05$). As expected from these data, a comparison of bimodal congruent and incongruent slopes yields highly significant results ($p < 0.01$), while there is no statistical difference between bimodal congruent and unimodal visual stimuli. At our representative central site, there are no increases in locking over time in any condition ($p > 0.05$). There are, however, condition differences regarding the magnitude of stimulus locking: although all conditions show significant stimulus locking ($p < 0.01$), peak correlations are significantly higher for bimodal congruent than incongruent and unimodal visual stimuli ($p < 0.01$). In summary, our results suggest two distinct multisensory interactions at work. At central recording sites we see a sensitivity to congruence between auditory and visual input streams, and at occipital sites a further time-dependent facilitation of visual stimulus-locking is evident that is suppressed when conflicting auditory information is present.

In the case of the spectrotemporal analysis approach, we first defined frequency bands of interest, guided by the visual stimulus locking results reported above (8-20 Hz and 20-30 Hz). The sum of spectrotemporal power in these bands was then correlated with stimulus speed using the shifting time window method just described, and peak correlation values (minimum in the case of the anti-correlation, and maximum for the correlation) were extracted from lag limits defined from the earlier results (200-300 ms and 0-150 ms, respectively). Finally, these time series were fitted using linear regression. No linear trend was seen for the 8-20 Hz band at occipital sites, but effect sizes were in the same range for bimodal congruent, incongruent, and unimodal visual conditions. In the case of the beta correlation, no linear trend was found for bimodal congruent and incongruent conditions at occipital sites ($p > 0.05$); however, locking of spectrotemporal power to unimodal visual stimuli was seen to increase over time in OZ ($p < 0.05$, $r^2 = 0.27$) and O1 (p <0.01, $r^2 = 0.47$).

4.5 Discussion

We measured EEG in 24 subjects while they viewed rotating Gabor patches and listened to frequency-modulated tones. Our results reveal two forms of stimulus locking of EEG to the temporal dynamics of visual stimuli, evident from changes in EEG power over time and from the temporal structure of the EEG waveform. Auditory stimuli presented alone do not lead to any measurable, structured changes in EEG power or waveform. However, analysis of bimodal trials shows that visual stimulus locking is modulated by the temporal congruence of simultaneous auditory input —locking of EEG waveforms to visual stimuli was reduced when the auditory input mismatched the temporal profile of the visual input. Furthermore, this multisensory interaction became even clearer when the time course of waveform locking was examined. Under congruent bimodal stimulation conditions, visual stimulus locking increased steadily over the duration of the trial, while the incongruent bimodal condition showed no such effect. Thus we suggest that stimulus locking is a suitable tool for studying and characterising the multisensory processing of dynamically changing auditory and visual stimuli.

In order to explore the importance of extended dynamic modulation of auditory and visual stimuli, we required a relatively long trial duration. This raised the need for a task to ensure participants' attention was maintained. We decided on a task that required attention to both stimulus components in parallel —to decide whether the auditory and visual signals matched —rather than directing the subject to focus on a single sensory modality. As a result, the task could not be directly transferred to single unimodal components of the audiovisual stimuli, and was instead applied to sequentially presented pairs of components from either modality. The simultaneous and sequential comparison of stimulus dynamics constitutes two different tasks and necessitates a careful comparison of unimodal and bimodal conditions. However, the task allows us to be confident of subjects' attentiveness during the experiment.

Here, as a first step, we report results for signals measured at electrodes, i.e. in sensor space. An EEG signal is a combination of signals from different cortical sources and noise. It would be desirable to further process the data in order to translate our findings into source space by performing a source localisation that would attempt to isolate the location and activity of the cortical sources involved. However, this requires additional assumptions and currently there is no available method that—is generally accepted and free of problems (Nunez 1981). An additional issue is that algorithms for source localization make assumptions about the interaction or independence of different signals.

Work is still under way to develop synchrony measures in source space (e.g. Marzetti et al. 2008). Although our results do not rely on a spatial interpretation, we do make some tentative assumptions here: that EEG measured at occipital channels captures activation in early visual areas proximal to these measurement sites, whereas effects we find at centro-parietal electrodes reflect cortical processing at some higher stages. A more detailed investigation of the cortical sources of the stimulus locking phenomena described here must be left to future work.

We did not see any evidence for stimulus locking to auditory stimuli, or to the auditory component of audiovisual stimuli. This is surprising, as it is commonly assumed that the temporal reliability of the auditory system is higher than the visual. As a consequence, its contribution in optimal sensory fusion is high for temporal estimates (e.g. Bresciani et al. 2008) and small for spatial estimates (e.g. Körding et al. 2007). In addition, the modulation of brain responses to changes in amplitude or frequency of simple auditory stimuli is a well-established phenomenon. Auditory entrainment has been optimally found in response to 40 Hz auditory stimulation (Galambos et al. 1981), but has also been reported at lower frequencies (e.g. Ding et al. 2006). A systematic examination of the steady-state response to sinusoidal frequency modulation by Picton et al. (1987) found that for lower modulation frequencies, responses were most reliable between 3 and 7 Hz. The amplitude and phase of responses to tones modulated at the rates of change used in our study were not found to be significantly reliable, and thus it may simply be the case that the auditory stimuli used here change too slowly. Indeed, preliminary results of current work with 3 Hz frequency modulation have shown a tendency to a locking of 40 Hz power to the dynamics of the auditory stimulus.

We applied two analysis approaches to investigate how the dynamics of the stimulus are reflected in the power and phase of the measured EEG signals, respectively. The first approach, spectrotemporal analysis, correlates power changes with stimulus speed, and can reveal locking of EEG power changes to stimulus dynamics at frequencies beyond the rate of change of the stimulus. The second approach, waveform analysis, correlates ERP waveforms with stimulus speed, and thus mirrors a stable phase-relationship between EEG signals and the time-course of the stimulus. Due to the use of correlation, the entrainment we observe for the EEG waveform must result from the phase-locking of frequencies within the range contained in the stimulus speed. Previously, locking of LFP power to dynamic visual stimulation has been investigated in cat visual cortex (Kayser & König 2004) using the same spectrotemporal locking analysis. In addition, entrainment or phase-locking of neural responses to regularly repeating stimuli has been extensively studied (e.g. Rager & Singer 1998) and this steady-state cortical response has been used to investigate many other phenomena, especially attention (e.g. Muller et al. 2003). Our analysis approaches are in line with both kinds of locking mechanisms, but are applied here to explore stimulus locking of human EEG to continuously, irregularly changing visual stimuli.

Our spectrotemporal approach investigated locking to all frequencies between 0 and 100 Hz. Within this range, different frequency bands have been defined, including delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (> 30 Hz) and these bands have been suggested to reflect different functionalities (see Steriade et al. (1990) for an early review). We found two instances of spectrotemporal locking that are centred at distinct frequencies (20-30 and 8-20 Hz) and furthermore differ in correlation lag and direction of effect. As such, we suggest that we are dealing with two distinct spectrotemporal locking processes with different functional significance.

The first effect found was that induced power changes in the beta (20-30 Hz) range lock to the speed profile of visual input with a lag of 92 ms. This is in agreement with previous results that found locking between 20-35 Hz at around 100 ms lag in cat visual cortex (Kayser & König 2004). Curiously, the frequency showing the strongest locking in our results is close to 25 Hz, which is exactly the frame rate at which our movies were created, but is far from the refresh rate of the monitor (100 Hz). However, it is not clear whether the correspondence with the frame rate is simply coincidental or not, which brings into question the functional significance of the 20-30 Hz band. There is a crucial difference between our study and the earlier work in cats —humans have a comparatively lower flicker fusion frequency than cat, so the 25 Hz frame rate can be considered to be adequate for human viewers. Although we have clearly found evidence for spectrotemporal locking to irregular visual stimulation, we cannot generalise this finding to more natural conditions.

The second effect revealed by the spectrotemporal anaylsis was an anti-correlation in a broad alpha band (8-20 Hz). The human occipital alpha rhythm has traditionally been associated with a state of cortical rest, as put forward by the idling hypothesis. More recent studies have extended the idling hypothesis, proposing that a strong alpha rhythm characterises a dominance of “top-down” processing in the absence of any external stimulation (von Stein et al. (2000), see Palva & Palva (2007) for a general review). Although the results here span a frequency band too broad to compare to the classical alpha band (8-13 Hz), our results fit with the idling hypothesis account, as stimulation strength anti-correlates with alpha power, with stronger input leading to reduced alpha oscillations. The anti-correlation found here does not show any modulation by auditory input, so we assume that the spectrotemporal locking we have observed is a purely visual phenomenon.

Overall, the effect sizes found using the spectrotemporal approach were small. These correlations are calculated using total power, which includes both induced (not phase-locked to stimulus onset) and evoked (phase-locked to stimulus onset) components (Tallon-Baudry & Bertrand 1999). Evoked power constitutes only a small part of total power —here approximately 15% —so if the stimulus-locked power changes are indeed consistent in their phase with respect to the stimulus dynamics, then the stimulus-locking effect may be hidden in the total power. In comparison, Kayser and König’s results were indeed larger, with correlations of approximately 0.13 compared to our 0.01, but were based on intra-cranial LFP measurements that do not suffer from the spatial smearing inherent in EEG.

The second approach we employed, waveform analysis, correlated ERP waveforms with stimulus speed. Our results revealed locking to visual stimulus dynamics, distributed across many measurement sites. An examination of the characteristics of the correlation results suggests at least two distinct locking mechanisms: one measured occipitally with a positive correlation at a lag of 78 ms, interpreted as an early visual process; and the other appearing at centro-parietal sites as positive locking at lags of -60 and 390 ms, interpreted as later stages of visual processing. A positive correlation indicates that the phase of the EEG waveform is aligned to the dynamics of the stimulus at a given time lag, however the relative magnitude of both signals also contributes to this measure of locking and we cannot isolate the role of phase alignment from the role of amplitude changes. As mentioned above, we can furthermore specify that entrained frequencies must lie in the range contained in the speed profiles of our stimuli (0.5-4 Hz). Although steady-state responses are typically evoked using stimuli presented at higher, regular stimulation frequencies, entrainment has also been observed for frequencies in the delta range (Ding et al. 2006). Less regular stimulus trains jittered in time within this frequency range have also been found to have an entrainment effect on visual and auditory cortex, which has been shown to be the result of an instantaneous phase reset of the ongoing oscillatory activity (Lakatos et al. 2008; 2005). Hence, we assume that in the case of our much more irregular stimuli, faster visual input leads to a stronger phase-reset across a large population of neurons, and thus to stronger EEG amplitude that is phase-aligned with the stimulus. Furthermore, we found that entrainment to visual stimuli increased over stimulus duration at occipital sites, indicating that phase alignment to our stimuli is a gradual process —there is no instantaneous phase reset.

The role of ongoing neural oscillations has been emphasised in multisensory research, in particular frequencies in the gamma range (see Senkowski et al. (2008) for a recent review). The entrainment of lower frequency oscillations, which are of particular interest here due to our choice of stimuli, has also been implicated in multisensory processes. In auditory cortex, the phase of all ongoing neuronal oscillations has been proposed to be reset by tactile or visual input, with this general phase reset hypothesised to allow more effective processing of subsequent auditory signals arriving at an optimal phase of this reset activity (Lakatos et al. 2007, Kayser et al. 2008). Such a mechanism has also been proposed to be useful for audiovisual speech, with the regularities of the visual input allowing a facilitation of temporally matched auditory input in noisy backgrounds (Schroeder et al. 2008). In conclusion, crossmodal resetting of cortical activity in a “unimodal” area may allow prioritised processing of temporally corresponding stimuli of the preferred modality, which would be of great importance for multisensory processing.

Our waveform analysis results for audiovisual conditions can be related to these findings. In general, we found stronger visual locking to congruent than incongruent audiovisual stimuli. This difference in entrainment to visual dynamics must be due to an effect of the auditory stimulus dynamics. One possible interpretation is that congruently timed auditory stimulation enhances the phase reset of visually entrained neurons. Furthermore, an examination of the time course of the visual entrainment of waveforms suggested two locking mechanisms, which seem to be differently involved in multisensory processing when the entrainment to purely visual stimuli is taken into account. The visual process measured at occipital sites shows a continual increase in visual entrainment over the duration of unimodal visual and congruent audiovisual stimulation, but a constant level of entrainment for mismatched audiovisual inputs. As such, incongruent auditory input seems to counteract visual entrainment, and the difference in entrainment between congruent and incongruent conditions unfolds gradually and slowly. The time course involved, in the magnitude of seconds, is a strong indication that this process is initially driven by bottom-up visual input, with information regarding multisensory congruence arriving later, most likely via feedback from higher areas. Along these lines, Noesselt et al. (2007) recently showed that modulations in primary unisensory areas in response to temporally aligned audiovisual patterns were a result of feedback from superior temporal sulcus. The visual locking found here at centroparietal sites shows no change over time, with congruent audiovisual input showing a clear advantage for locking over both incongruent and unimodal visual stimulation. Thus, this process depends on both input streams; however it is unclear whether this mechanism reflects an evaluation of congruence or is itself the result of a match/mismatch evaluation from elsewhere. Source analysis methods, and a direct evaluation of the directional interaction between the two waveform locking mechanisms, may help elucidate this in future work.

The behaviour of a dynamical system can be characterised by examining the relationship between its output and a known input. Often-used inputs are of two different kinds: impulses or continuous stimuli. In the case of a linear system, these two inputs yield equivalent results. To date, sensory processing has been analysed with an emphasis on this impulse-response concept, specifically multimodal interaction (see Calvert & Thesen (2004) for a general review of methodological approaches in multisensory research). Here we want to fully exploit the high temporal resolution of EEG to examine the nature of a temporal process, and it is rather the second kind of input which is of interest. The present analysis amounts to an investigation of the linearity of the audiovisual interactions. Indeed, the observation of a time-dependant increase in coupling in the bimodal congruent condition and unimodal visual condition is difficult to reconcile with a purely linear system. The characteristics of the stimuli and the analysis methods used in this study are better suited to such investigations regarding non-linear system behaviour than traditional impulse response approaches.

The integration of two sensory sources critically depends on the compatibility of the temporal dynamics of both information streams. We used stimulus locking to extended stimuli with complex temporal profiles as a tool to investigate multisensory interactions. Spectrotemporal locking revealed a bottom-up, purely visual effect. In contrast, waveform locking showed modulation by auditory congruency and a differential effect over time. We thus propose that spectrotemporal and waveform locking reflect different mechanisms involved in the processing of dynamic audiovisual stimuli, and that these analysis approaches may prove useful in future research.

5

A Behavioral Approach: Crossmodal Integration for the Control of Overt Attention*

5.1 Context

In everyday life, our brains decide about the relevance of huge amounts of sensory input. Further complicating this situation, the input is distributed over different modalities. This raises the question of how different sources of information interact for the control of overt attention during free exploration of the environment under natural conditions.

Previous Chapters dealt with the question of how the presence of different sources of information are integrated at the cortical level. In the first Chapter we were interested in the integration of binocular information. The second Chapter dealt with the question of how cortical connectivity underpins the integration of spatially separated visual information. In the third Chapter we extended this question to different modalities and investigated how signals originating from different modalities are integrated in the case of dynamic input. During these Chapters we ignored to a large extent the behavioral component of these underlying neuronal processes. In this Chapter we investigate how behavior is influenced by the presence of different sources of input. The integration we observed at the cortical level leaves open the question of how the behavior is influenced. In the behavioral level, different modalities may work independently or interact to determine the resulting motor output.

In order to investigate this question, we focused on human overt attention by means of eye tracking methodology. To collect information from the environment, humans extensively use their visual system by deploying their attentional resources to different regions of a scene. During the exploration of a scene, fast eye movements called saccades intermingling with relatively stable fixation periods form the basis of this process. By measuring the density of the fixated points, one can obtain an empirical estimate of the

---

*This chapter is based on a publication Onat et al. (2007)

saliency of different regions on an image. Such empirical saliency maps can later be used in order to evaluate different integration mechanisms.

We presented natural images and lateralized natural sounds in a variety of conditions, and measured the eye movements of human subjects. We show that, in multimodal conditions, fixation probabilities increase on the side of the image where the sound originates showing that, at a coarser scale, lateralized auditory stimulation topographically increases the salience of the visual field. However, this shift of attention is specific because the probability of fixation of a given location on the side of the sound scales with the saliency of the visual stimulus, meaning that the selection of fixation points during multimodal conditions is dependent on the saliencies of both auditory and visual stimuli. Further analysis shows that a linear combination of both unimodal saliencies provides a good model for this integration process, which is optimal according to information theoretic criteria.

Our results support a functional joint saliency map, which integrates different unimodal saliencies before any decision is taken about the subsequent fixation point. These results provide guidelines for the performance and architecture of any model of overt attention that deals with more than one modality.

5.2 Introduction

How are different sources of information integrated in the brain while we overtly explore natural multimodal scenes? It is well established that the speed and accuracy of eye movements in performance tasks improve significantly with congruent multimodal stimulation (Corneil & Munoz 1996, Corneil et al. 2002, Arndt & Colonius 2003). This supports the claim that sensory evidence is integrated before a motor response. Indeed, recent findings indicate that areas in the brain may interact in many different ways (Driver & Spence 2000, Macaluso & Driver 2005). The convergence of unimodal information creates multimodal functionality (Meredith & Stein 1986, Beauchamp et al. 2004) even at low-level areas traditionally conceived as unimodal (Macaluso et al. 2000, Calvert et al. 1997, Ghazanfar et al. 2005); evidence is also currently mounting for early feedforward convergence of unimodal signals (Molholm et al. 2002, Fu et al. 2003, Foxe & Schroeder 2005, Kayser et al. 2005). Little is known, however, about integration processes under the relevant operational —i.e. natural —conditions. Most importantly, we lack a formal description of the integration process during overt attention. It is important to develop formalizations using behavioral data as it reflects the final outcome of the processes within the CNS.

Current models of overt attention are based on the concept of a saliency map: A given stimulus is first separated into different feature channels; after the local contrast within each feature space is computed, these channels are then combined, possibly incorporating task-specific biases (Koch & Ullman 1985, Itti & Koch 2001, Parkhurst et al. 2002, Peters et al. 2005). The selection of the next fixation point from a saliency map involves a strong non-linearity. This is typically implemented as a winner-takes-all mechanism. The timing of this non-linearity with respect to the integration of multiple feature channels crucially influences both the performance and structure of the resulting system. In principle, three idealized multimodal integration schemes can be considered.

Early interaction —The information from different modalities could be integrated early, before the computation of a saliency measure and the selection of a fixation point. Indeed, many studies provide evidence for an interaction between signals in early sensory cortices (Calvert et al. 2000, Kayser et al. 2005). An integration in the form of interaction may be the result of modulatory effects of extra-modal signals during computation of saliency maps within a given modality. Or alternatively such an interaction may be caused by multiplicative integration of unimodal saliencies after the saliency maps for each modality are computed. The selection of the most salient point after crossmodal interaction has taken place imposes an expansive nonlinearity. Consequently, sensory signals from different modalities should interact supralinearly for the control of gaze movements.

Linear integration —Alternatively, saliency could be computed separately for different modalities and subsequently be combined linearly before fixation selection. Recent research shows that the bulk of neuronal operation underlying multisensory integration in superior colliculus can be well described by a summation of unimodal channels (Stanford et al. 2005). Within this scheme multimodal saliency would be the result of the linear combination of unimodal saliencies. From an information-theoretic perspective, this type of linear summation is optimal, as is the resulting multimodal map, in the sense that the final information gain is equal to the sums of the unimodal information gains.

Late combination —No true integration between modalities occurs. Instead, overt behavior results from competition between candidate fixation points in the independent unimodal saliency maps. The implementation of such a max operator results in a sublinear integration of the unimodal saliency maps. Although improvements in saccade latencies and fixation accuracies in (non-natural) multimodal stimulus conditions have been used to support the counter-claim that crossmodal integration does take place (Corneil et al. 2002, Arndt & Colonius 2003), this hypothesis still warrants investigation under natural free-viewing conditions.

We presented human subjects with lateralized natural sounds and natural images in a variety of conditions and tracked their eye movements as a measure of overt attentional allocation. These measurements were used to compute empirically determined saliency maps, thus allowing us to investigate the above hypotheses.

5.3 Materials & Methods

5.3.1 Participants and recording

Forty-two subjects (19 males, mean age 23) participated in the experiment. All subjects gave informed written consent but were naive to the purpose of the experiment. All experimental procedures were in compliance with guidelines described in Declaration of Helsinki. Each subject performed only one session.

The experiments were conducted in a small dimly lit room. The subjects sat in a stable chair with back support, facing a monitor. A video-based head-mounted eye tracker (Eye Link 2, SR Research, Ontario, Canada) with sampling rate of 250 Hz and nominal spatial resolution of 0.01° was used for recording eye movements.

For calibration purposes, subjects were asked to fixate points appearing in a random sequence on a 3 × 3 grid by using built-in programs provided with Eye Link (similar to Tatler et al. (2006)). This procedure was repeated several times to obtain optimal accuracy of calibration. It lasted for several minutes, thus allowing subjects to adapt to the conditions in the experimental room. All the data analyzed in the present article were obtained from recordings with an average absolute global error of less then 0.3°.

During the experiment, a fixation point appeared in the center of the screen before each stimulus presentation. The experimenter triggered the stimulus display only after the subject had fixated this point. The data obtained during this control fixation were used to correct for slow drifts of the eye tracker; that is, if drift errors were high a new calibration protocol was started again. Subjects could take a break and remove the headset at any time. In those instances, which occurred rarely, the continuation of the experiment began with the calibration procedure described above.

5.3.2 Stimuli

The multimodal stimuli consisted of images and sounds. Images depicted natural scenes like forests, bushes, branches, hills, open landscapes, close ups of grasses and stones, but also to a limited extent human artifacts, for example, roads, house parts. Some of these stimuli were used in a previous study (Einhäuser & König 2003). The photographs were taken using a 3.3 megapixel digital camera (Nikon Coolpix 995, Tokyo, Japan), downsampled to a resolution of 1024 ×768 pixels, and converted to grayscale. The images were displayed on a 21-in. CRT monitor (SyncMaster 1100DF, Samsung Electronics, Suwon, South Korea), at a resolution of 1024 ×768 pixels and with a refresh rate of 120 Hz. The distance of the monitor from the subject's eyes was 80 cm. The stimuli covered $28^\circ \times 21^\circ$ of visual angle on the horizontal and vertical axes, respectively. The screen was calibrated for optimal contrast and brightness using a commercial colormeter (EyeOne Display, GretagMacBeth, Regensburg, Switzerland). The natural auditory stimuli were taken from free samples of a commercial Internet Source (Askland Technologies Inc., Victorville, Canada). Overall, 32 different sound tracks were used. All of these were songs of different birds, thus in accordance with the semantic content of the images presented. The auditory stimuli were generated using a sound card (Sound Blaster Audigy EX ZE Platinium Pro, Creative Labs, Singapore) and loudspeakers (Logitech 2.1 Z-3, CA, USA). Loudspeakers flanked both sides of the monitor at a distance of 20 cm, at the same depth plane as the screen. In order to avoid the speakers attracting the attention of the subjects, they were hidden behind black curtains. Sounds were played from the left or right speaker, depending on the experimental condition. The auditory signal amplitude was in the range of 50-70 and 55-75 dB for left and right conditions, respectively. The slight increase was due to the acoustic conditions in the experimental room.

5.3.3 Experimental paradigm

We employed two unimodal conditions (visual and auditory) and one multimodal condition (audiovisual) (5.1A). During the visual condition (V), 32 natural images were presented. The auditory condition used 16×2 presentations of natural sounds, originating from left (AL) and right (AR) side relative to the subject's visual field.

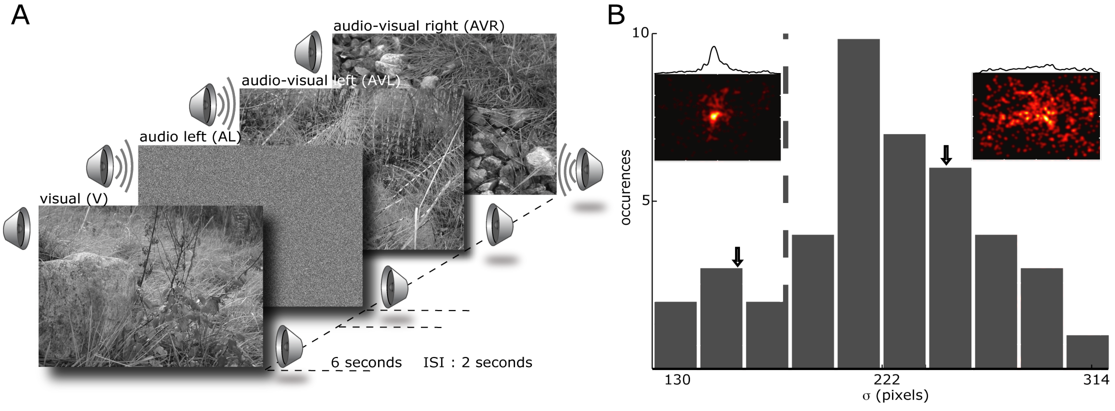

Figure 5.1: **Experimental Paradigm and Subject Selection.** **(A)** Five different experimental conditions are used to analyze the effect of unimodal (visual and auditory) and crossmodal (audiovisual) stimuli on overt behavior. 42 subjects studied 32 natural images in three different conditions. In the visual condition (V), natural images were presented. In multimodal conditions, the images were paired with localized natural sounds originating from either the left (AVL) or right (AVR) side of the monitor. In the remaining conditions (AL and AR), the sole effect of the localized auditory stimulus was characterized. The duration of stimulus presentation in all conditions was 6 seconds. **(B)** Distribution of values of each subject's σ, which characterizes the horizontal spread of the pdf $p_{s,v}(x, y)$ of a given subject's at condition V. $p_{s,v}(x, y)$ from two subjects are shown in the inset with corresponding horizontal marginal distributions shown above. Arrows mark the histogram bins to which these subjects belong. The vertical dashed line indicates the threshold σ value required for a given subject to be included in further analysis.

During auditory conditions, we presented auditory stimuli jointly with white noise images. These were constructed by shuffling the pixel coordinates of the original images. They lack any spatial structure and as a result do not bias the fixation behavior of the subjects. Obtaining a truly unimodal saliency map for auditory conditions adds some undesirable technical issues. Firstly, due to the operation of the eye tracker and monitor truly zero-light conditions are hard to achieve. Secondly, presenting a dark stimulus leads to more fixations outside the dynamics range of the monitor and eye tracker. Finally, a sudden drastic change in mean luminance of the visual stimulus would introduce non-stationarities in the form of dark adaptation and create a potential confound. To avoid these problems, we presented white noise images of identical mean luminance as the natural pictures.

The multimodal conditions (AVL and AVR) each comprised the simultaneous presentation of 32 auditory and visual stimuli pairs, without any onset or offset asynchrony. In order to balance the stimulus set, a new pairing of audiovisual stimuli was presented on each side to each subject. Stimuli were shown in a pseudorandom order, with a different permutation used for each subject. Each stimulus was presented for 6 s. A given session contained 128 stimulus presentations and lasted in total for up to 50 min. The only instruction given to the subjects was to watch and listen carefully to the images and sounds. No information about the presence of the speakers at both sides of the monitor or the lateralization of the auditory stimuli was provided to the subjects.

5.3.4 Data analysis

We defined fixation points and intervening saccades using a set of heuristics. A saccade was characterized by an acceleration exceeding 8000 deg/s^2, a velocity above 30 deg/s, a motion threshold of 0.1°, and a duration of more than 4 ms. The intervening episodes were defined as fixation events. The result of applying these parameters was plotted and was visually assessed to check that they produce reasonable results.

5.3.5 Probability distributions

From the fixation points of the individual subjects, we built probability density functions (pdf). The first fixation on each stimulus was discarded, as it was always located at the position of the preceding fixation cross in the center of the image. A pdf, $p_{s,i,c}(x,y)$, for a given subject s, image i, and condition c was calculated as in equation 5.1.

$$p_{s,i,c}(x,y,t) = \frac{1}{F}\sum_{f=1}^{F}\delta(x-x_f)\delta(y-y_f)\delta(t-t_f) \tag{5.1}$$

with $\delta(x)$ as the discrete Dirac function (the Dirac function is equal to zero unless its argument is zero, and it has a unit integral). x_f, y_f, and t_f are the coordinates and time of the f$^{\text{th}}$ fixation. F is the total number of fixations. We distinguish three different pdfs for a given condition with respect to how these individual pdfs were averaged: subject, image, and spatio-temporal pdfs. Subject pdfs $p_{s,c}(x,y)$ for a given subject s and condition c were built by averaging all the pdfs obtained from a given subject over the images, without mixing the conditions, according to 5.2

$$p_{s,c}(x,y) = \frac{1}{I}\sum_{i}^{I} p_{s,i,c}(x,y) \tag{5.2}$$

Image pdfs ($p(x,y)$) and spatio-temporal pdfs ($p(x,t)$) were similarly computed by averaging over the appropriate dimensions. Image pdfs inform us about consistent biases that influence subjects' gaze and are therefore empirically determined saliency maps specific to a given image. Raw pdfs are matrices of the same size as an image and store fixation counts in their entries. In all cases, these raw pdfs were smoothed for further analysis by convolution. A circular two-dimensional Gaussian kernel was used, which had a unit integral and a width parameter σ with a value of 0.6° unless otherwise stated. This spatial scale is twice as large as the maximal calibration error and maintains sufficient spatial structure for data analysis.

5.3.6 PDF parameterization

Several parameters were extracted from the pdfs,

$$\mu_c = \frac{\sum_x p_c(x)x}{\sum_x p_c(x)} \tag{5.3}$$

The center of gravity is measured according to Equation 5.3 in order to quantify the global shift of subject and image pdfs. μ_c is the center of gravity along the X-axis for condition c, $p_c(x)$ is the marginal distribution of a pdf along the X-axis. In condition V, fixation probabilities were usually distributed symmetrically on both sides of the visual field with centralized center of gravity values. This simple statistic successfully quantifies any bias toward the sound location,

$$\sigma = [\sum_x (x - \mu_V)^2 p_V(x)]^{1/2} \tag{5.4}$$

The spread of a given pdf was measured from a subject pdf under condition V using Equation 5.4 in order to quantify how explorative that subject's scanning behavior was. σ is the spread along the X-axis, $p_V(x)$ is the marginal distribution along the X-axis of the subject pdf, and μ_V is the center of gravity computed as in Equation 5.3. The marginal distributions arising from condition V were well-behaved (5.1 inset), thus allowing us to examine the explorative behavior of subjects. Seven of 42 subjects did not engage in explorative viewing during the analysis of the images, resulting in small spread values. These subjects were excluded from further analysis.

The spatio-temporal pdfs, $p_c(x,t)$, contain information about how the fixation density along the horizontal axis varies over time. We assessed the statistical differences of pdfs from different conditions in a time-localized manner; that is, we obtained a p-value as a function of time for three pairs of conditions (AVL and V; AVR and V; AL and AR). A two-sided Kolmogorov-Smirnov goodness-of-fit hypothesis test in corresponding temporal portions of the pdfs was used with significance level α set to .001. This was done after binning the probability distribution $p_c(x,t)$ over the time axis with a bin size of 240 ms, yielding 25 time intervals for comparison per pair of conditions. A temporal interval over which the null hypothesis was rejected in at least in two of the condition pairs was considered as a temporal interval of interest.

The similarity between image pdfs obtained from the same image under different conditions —for example, $p_{i,V}(x,y)$ of image i and condition V and $p_{i,AV}(x,y)$ of the same image and AV conditions (i.e., either AVL or AVR) —was evaluated by computing r^2_{VAV}. Before the coefficients were calculated, the AV image pdfs were first normalized to $p^N_{i,AV}$ according to Equation 5.5

$$p^N_{i,AV} = \frac{< p_V(x) >_i}{< p_{AV}(x) >_i} p_{i,AV}(x,y) \tag{5.5}$$

This normalization corrects the image pdfs of a given bimodal condition for the global effect of the sound location, so that the expected number of fixations over the horizontal axis is the same over different conditions. The resulting distribution of $r^2_{V,AV}$ values was then compared to a control distribution of r^2 values, which measure the baseline correlation between image pdfs coming from the same condition pairs (i.e., V and AVL, or V and AVR) but differing images. The Kullback-Leibler (KL) divergence, which does not assume any a priori relationship between two distributions, was used to quantify the similarity of the different image pdfs obtained from different conditions. This measure was evaluated according to the following formula 5.6, where $D_{KL}(p_{i,c_1}, p_{i,c_2})$ denotes the KL divergence measure between two pdfs, $p_{i,c_1}(x,y), p_{i,c_2}(x,y)$ in bits,

$$D_{KL}[p_{i,c_1}(x,y), p_{i,c_2}(x,y)] = H[p_{i,c_1}(x,y), p_{i,c_2}(x,y)] - H[p_{i,c1}(x,y)] \tag{5.6}$$

The KL divergence measures the difference between the cross-entropy of two probability distributions and the entropy of one of them. The cross-entropy is always greater than or equal to the entropy; therefore, the KL divergence is always greater than or equal to zero, which allows its usage as a distance measure between two different pdfs. However, unlike other distance measurements, it is not symmetric. Therefore, it is used to measure the distance between the prior and posterior distributions. In comparing different image pdfs, we used the condition V as the prior distribution. One problem we encountered was zero entries in the probability distributions. As the logarithm of zero is not defined, a small constant ($c = 10^{-9}$) was added to all entries in the pdf. The precise choice of this constant did not make a difference to the results of our analysis.

5.3.7 Modeling the cross-modal interaction

In order to quantify the cross-modal interaction, we carried out a multiple regression analysis. We devised a model (Equation 5.7) with a cross-product interaction term using smoothed unimodal and multimodal pdfs as independent and dependent variables, respectively,

$$p_{i,AV} = \beta_{i,0} + \beta_{i,1} \cdot_{i,V} + \beta_{i,2} \cdot p_{i,A} + \beta_{i,3} \cdot p_{i,V-A} \tag{5.7}$$

In Equation 5.7, $p_{i,AV}$, $p_{i,V}$, and $p_{i,A}$ are the image pdfs of image i at audiovisual, visual, and auditory conditions, respectively. The interaction term $p_{i,VA}$ is supposed to approximate the image pdf that would arise from a multiplicative cross-modal interaction. It is created by the element-wise multiplication of both unimodal image pdfs and renormalized to a unit integral.

The integrative process was further characterized by constructing integration plots. The probability of fixation at each x-y location was extracted from the 32 image pdfs of visual, auditory, and bimodal conditions, yielding a triplet of values representing the saliency. A given triplet defines the saliency of a given image location in the multimodal condition as a function of the saliency of the same location in both unimodal conditions, represented by the point $(p_V(x,y), p_A(x,y), p_{AV}(x,y))$ in the integration plot. These points were irregularly distributed and filled the three-dimensional space unevenly. We discarded the 15% of the values which lay in sparsely distributed regions, and we concentrated instead on the region where most of the data were located. The data points inside this region of interest were then binned, yielding an expected value and variance for each bin. Weighted least square analysis was carried out to approximate the distribution by estimating the coefficients of the following equation:

$$\frac{p_{AV}}{g(p_{AV})} = \beta_0 + \beta_1 \cdot \frac{p_V}{g(p_V)} + \beta_2 \cdot \frac{p_A}{g(p_A)} + \beta_3 \cdot \frac{p_{V \cdot A}}{g(p_{V \cdot A})} \tag{5.8}$$

The difference between the above equation and Equation 5.7 is that Equation 5.8 does not take different images into consideration and pools the data over images and space. Additionally, each individual probability value is normalized by its geometric mean ($g(pc)$), which normalizes for its individual range thus allowing a direct comparison of the regression coefficients.

5.3.8 Luminance contrast measurements

LC was computed as the standard deviation of the luminance values of the pixels inside a square patch of about 1° centered at fixation positions. The luminance contrast computed at the fixation points over images and subjects yielded the actual contrast distribution. This distribution was compared to a control distribution to evaluate a potential bias at the fixation points. An unbiased pool of fixation coordinates served as the control distribution —for a given image, this was constructed by taking all fixations from all images other than the image under consideration. This control distribution takes the center bias of the subjects' fixations into account, as well as any potential systematic effect in our stimulus database (Baddeley & Tatler 2006, Tatler et al. 2005). The contrast effect at fixation points was computed by taking the ratio of the average contrast values at control and actual fixations. In order to evaluate the luminance contrast effect over time, the actual and control fixation points were separated according to their occurrences in time. The analysis was carried out using different temporal bin sizes ranging from 100 to 1000 ms. All analysis was carried out using MatLab (Mathworks, Natick, MA, USA).

5.4 Results

First, we analyze the effect of the lateralized auditory stimuli on the fixation behavior of subjects during the study of natural images Fig. 5.1A. Second, we characterize the temporal interval during which the influence of auditory stimuli is strongest. Third, we demonstrate the specific interaction of visual and auditory information. And finally, we address the predictions derived from the three hypotheses of crossmodal interaction stated above.

5.4.1 Subjects' Gaze Is Biased towards the Sound Location.

In Fig. 5.2, the spatial distribution of fixation points averaged over images in all 5 conditions is shown for 2 subjects. In the visual condition (V), the fixation density covers a large area and is equally distributed over the left and right half of the screen, with neither of the subjects showing a consistent horizontal bias. The center of gravity (μ_V) is located at 505 and 541 pixels respectively (white crosses) for these two subjects, in the close vicinity of the center of the screen (located at 512 pixels). In the multimodal conditions (AVL and AVR), both subjects show a change in their fixation behavior. The horizontal distance between μ_{AVL} and μ_{AVR} is 221 and 90 pixels for the two subjects respectively. Thus, in these two subjects, combined visual and auditory stimulation introduces a robust bias of fixation towards the side of sound presentation.

Auditory unimodal conditions (AL and AR) induce different patterns of fixation (Fig. 5.2 right hand columns, note different scales of color bars). First, despite lateralized sound presentation, a non-negligible proportion of fixations are located close to the center. Nevertheless the lateralized sound stimulus induces a shift toward the sound location, even in the absence of any structured, meaningful visual stimulation. In most cases, the shift had an upwards component. Furthermore, the off-center fixations are less homogeneously distributed and their distribution does not qualitatively resemble the distribution of fixations in the visual condition.

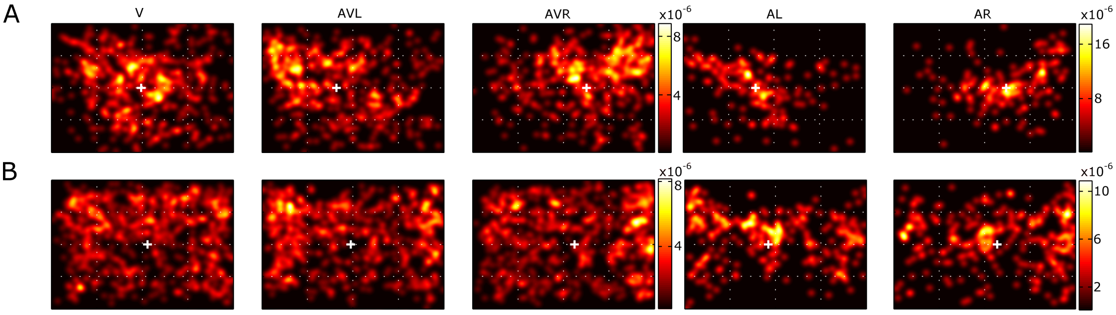

Figure 5.2: **Bias of Subjects' Gaze towards the Sound Location.** **(A, B)** $p_{s,c}(x,y)$ for two subjects in each condition. Each colorbar shows the scale of the pdf images located to its left; notice the differences in scale. White crosses denote the center of gravity of a given pdf along the horizontal axis. These pdfs were generated by convolving the original pdfs with a Gaussian kernel ($\sigma = 0.6°$).

The complete statistics of all subjects and images are shown in Fig. 5.3. The distribution of center of gravity shifts ($\mu_{AVL} - \mu_V$ and $\mu_{AVR} - \mu_V$) for all subjects over all images is skewed (Fig. 5.3). In the majority of subjects, we observe a moderate effect of lateralized sound presentation. A small group of subjects show only a small influence; one subject an extreme effect. The medians of the two distributions are both significantly different from zero (sign test, $p > 10^{-5}$). Crosses indicate the positions of the two example subjects described above. They represent the 70^{th} and 90^{th} percentiles of the distributions. Hence, the complete statistics support the observations reported for the two subjects above.

An analysis of the influence of auditory stimuli on the selection of fixation points in individual images (over all subjects) is shown in Fig. 5.3B. We see that for all visual stimuli the shift in average horizontal position of fixation points is towards the sound location (t-test, $p > 10^{-5}$). In both of the panels A and B, the distributions flank both sides of zero, with mean values of -37 and 36 pixels for $\mu_{AVL} - \mu_V$ and $\mu_{AVR} - \mu_V$ respectively. Thus, auditory stimulation introduces a robust bias of fixation towards the side of sound presentation for all natural visual stimuli investigated.

5.4.2 Effect is More Prominent in the First Half of Presentation

Next we analyze the temporal evolution of the above-described horizontal effect of the auditory stimulus. Fig. 5.4 depicts the difference between the spatio-temporal fixation probability density functions (pdf) for AVR and V, AVL and V, and AL and AR conditions. Comparing visual and multimodal conditions shortly after stimulus onset, (Fig. 5.4A, B lower plots) we observe a rapid shift in fixation density on the side of the auditory stimulus, which is sustained for the presentation period. This increase in fixation density is realized at the expense of fixations in the central region, and less so at the expense of fixations on the contra-lateral side. This can be seen by comparing the marginal distributions (averaged over time) originating from the pdfs used to calculate

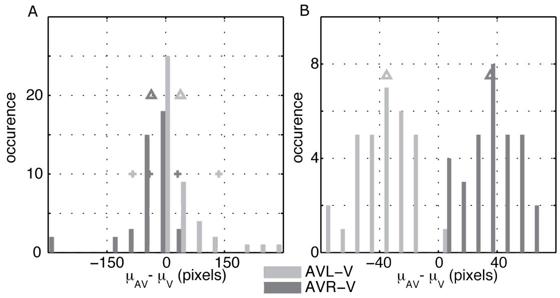

Figure 5.3: **Distribution of Fixation Density Shifts towards Sound Location.** **(A)** Centers of gravity (μ_c for condition c) were calculated for the fixation pdfs of each subject averaged over all images for multimodal conditions (AVR and AVL) and unimodal visual condition (V). The distributions of distances between multimodal and unimodal centers of gravity are shown. The averages are marked with arrowheads and equal -44 pixels for $\mu_{AVL} - \mu_V$ (*dark gray*) and 42 pixels for $\mu_{AVR} - \mu_V$ (*light gray*). The two distributions are statistically different (sign test, $p < 10^{-5}$). The plus signs mark the effect size for the subjects depicted in Fig. 5.2. **(B)** Shows the same measurement, this time calculated for fixation pdfs of each image averaged over all subjects. The two distributions are statistically different (t-test, $p < 10^{-5}$). The average values of the distributions are same as in (A). However, the variance of the shifts of gravity centers is bigger on subject pdfs compared to the image pdfs therefore resulting in different scales on the abscissa.

the difference (Fig. 5.4A, B upper plots). The intervals of time for which the differences reach significance level (two-sided KS-test, p=0.001) are indicated by vertical black bars. Comparing conditions AL and AR (Fig. 5.4 right panel), we observe an increase in fixation probability on the half of the horizontal axis corresponding to the side of the sound location. This difference decays only slowly over time.

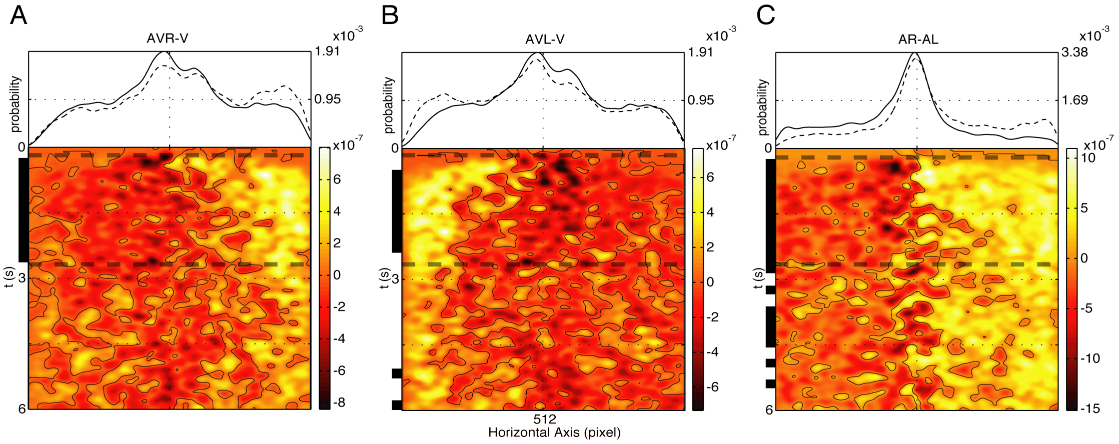

Figure 5.4: **Effect of Localized Auditory Stimulus is More Prominent in the First Half of Presentation.** The uppermost plots show two marginal probability distributions for the following pairs of conditions: **(A)** AVR and V, **(B)** AVL and V, **(C)** AR and AL, *dashed* and *solid* lines respectively. The lower plots depict the difference between the spatio-temporal pdfs of the same pairs of conditions. Contour lines are drawn at zero values. Along the times axis, the *black horizontal* lines mark the regions where the difference between the two pdfs is significant (two-sided KS-test, $\alpha = 0.001$). The *vertical dashed* line limits the temporal interval of interest, which is used for further analysis. The pdfs were generated using a Gaussian kernel ($\sigma = 0.4°$ and 70 ms).

5.4.3 Specific Interaction of Visual and Auditory Information

As a next step we investigate the integration of auditory and visual information in the temporal interval of interest. The first column of Fig. 5.5 depicts examples of the natural images used in the experiment. The other columns show the image pdfs computed over all subjects in conditions V, AVL and AVR. As these pdfs were computed over many subjects' fixations, they constitute empirically determined saliency maps. For each image, we observe a characteristic distribution of salient regions, i.e. regions with high fixation probabilities. It is important to note that the fixation densities are highly unevenly distributed, suggesting a similarity between subjects' behaviors. We computed the correlation coefficients between image pdfs generated from two subsets of 5 randomly selected subjects (repeating the same analysis 300 times). For all three conditions the distribution of coefficients over all images and repetitions peaked at around 0.6. This suggests that different subjects had similar behaviors for scrutinizing the image during the temporal interval of interest (240 - 2640 ms). It is not clear whether this was the result of the specific image content or shared search strategies between subjects.

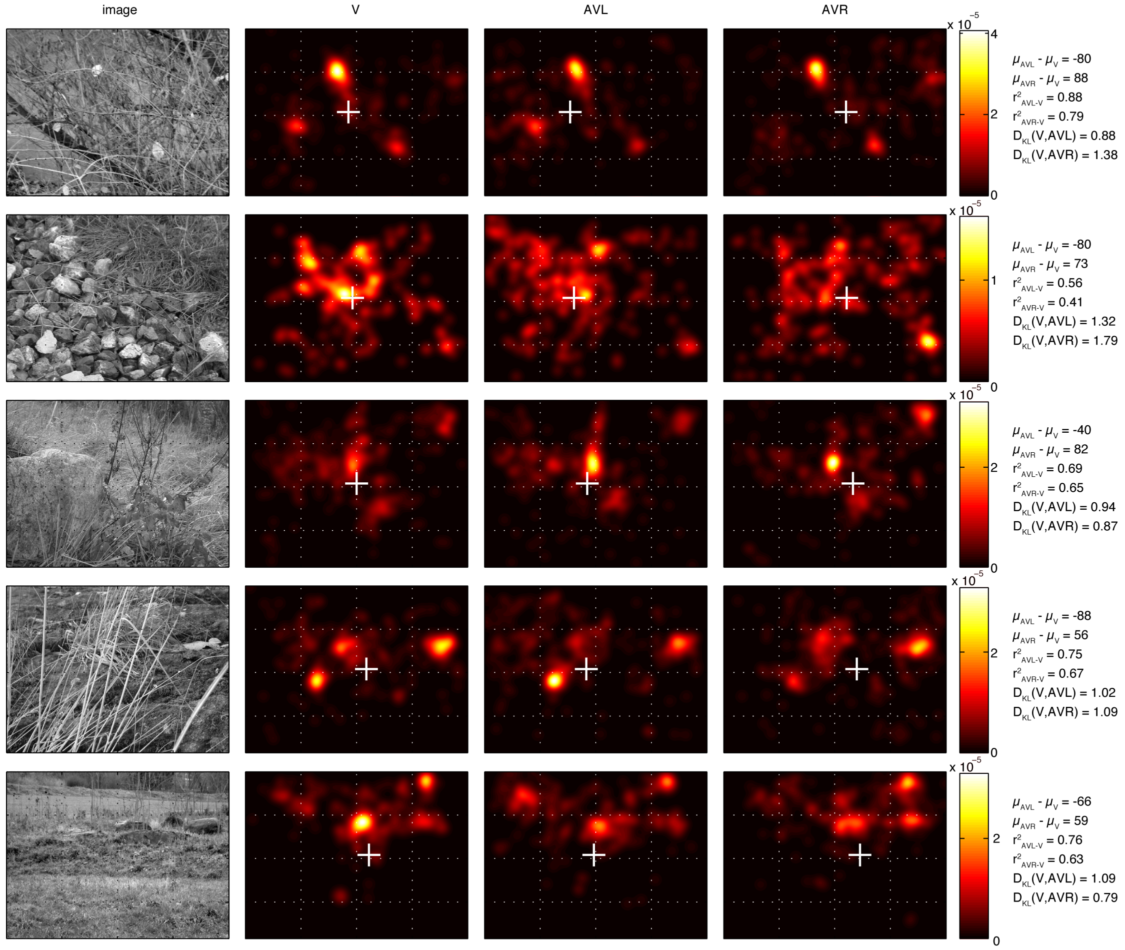

Figure 5.5: **Specific Interaction of Visual and Auditory Information.** Fixation pdfs,$p_{i,c}(x, y)$, for a given image i and conditions c (V, AVL and AVR) are shown, along with the corresponding natural image i. Each pdf constitutes an empirically determined saliency map. The saliency maps for each image are shown along rows, for each condition along columns. *White crosses* in each panel depict the centers of gravity of the pdfs. In multimodal conditions, the center of gravity shifts toward the side of auditory stimulation. Interestingly however, moving across each row we see that the salient spots for each image are conserved across conditions as shown by the high r^2 and low KL divergence values (right of colorbar). Fixation pdfs are computed inside the temporal interval of interest and convolved with a Gaussian kernel ($\sigma = 0.6°$).

The two right-hand columns show the respective distributions obtained in multimodal conditions. First, as noted earlier, the lateralized stimulus causes the center of gravity to shift along the horizontal axis (Fig. 5.5, white crosses). Importantly, the spatial distributions of the spots are alike in different conditions, but differ across images. The regions with high fixation probability in one multimodal condition (Fig. 5.5) still effectively attract gaze when the side of auditory stimulation is switched, as well as in the unimodal condition.

This observation is quantified by measuring the KL divergence and r^2 statistic between saliency maps (such as those in Fig. 5.5) belonging to unimodal and cross-modal conditions. For the examples shown in Fig. 5.5 we obtain a KL-divergence of 0.88, 1.32, 0.94, 1.02, 1.09 bits between V and AVL conditions and 1.38, 1.79, 0.87, 1.09, 0.79 bits between V and AVR. r^2 statistics range from 0.4 to 0.97, indicating that a substantial part of the total variance of the multimodal conditions is explained by the distribution of fixation points in the unimodal visual condition.

The distribution of KL-divergence values obtained from the complete dataset (32 images times 2 auditory stimulus locations) is presented in 5.6. The more similar the two pdfs are, the closer the KL-divergence values get to zero; zero being the lower limit in the case of identity. This distribution is centered at 1.08 ± 0.03 bits (mean ±SEM). This is significantly different to the mean of the control distribution (3.45 ± 0.02 bits), which was created using 3200 randomly selected non-matched V-AV pairs. The control distribution provides the upper limit for KL-divergence values given our data set. Hence, given the distribution of fixation points on an image in the visual condition, the amount of information necessary to describe the distribution of fixation points on this image in the multimodal conditions is about one third of the information necessary to describe the difference in fixation points on different images in these conditions.

These results are supported by a more conventional linear measure. Fig. 5.6B shows the distribution of actual r^2 values calculated between image pdfs from multimodal and unimodal conditions originating from the same image. The distribution is centered at 0.71 ± 0.13 and the difference between this measure and a control r^2 measure calculated from shuffled image pairs is highly significant (t-test, $p > 10^{-5}$). This implies that for the majority of images, the unimodal fixation pdfs account for more than half of the variance in the observed distribution of fixation points in multimodal conditions. Hence, the bias of gaze movements towards the side of the auditory stimulus largely conserves the characteristics of the visual saliency distribution. Therefore the behavior of the subjects under the simultaneous presence of auditory and visual stimuli is an integration of both modalities.

As a complementary approach, we investigate the effect of multimodal stimuli on the relationship between visual stimulus properties and the selection of fixation points. Several previous studies investigating human eye movements under natural conditions describe a systematic increase of luminance contrast at fixation points (Reinagel & Zador 1999, Tatler et al. 2005). If the auditory stimuli cause an orientation behavior independent of the visual stimuli, then we can expect the luminance contrast at fixation points to be reduced. If a true integration occurs, we expect this correlation between luminance contrast and probability of fixation to be maintained under multimodal stimulus conditions. Fig. 5.6C shows the ratio of luminance contrast at actual fixations and control locations for the unimodal and both multimodal conditions. Nearly all values ($\log \frac{actual}{control}$) are greater than zero, indicating a positive correlation between fixation points and lumi-

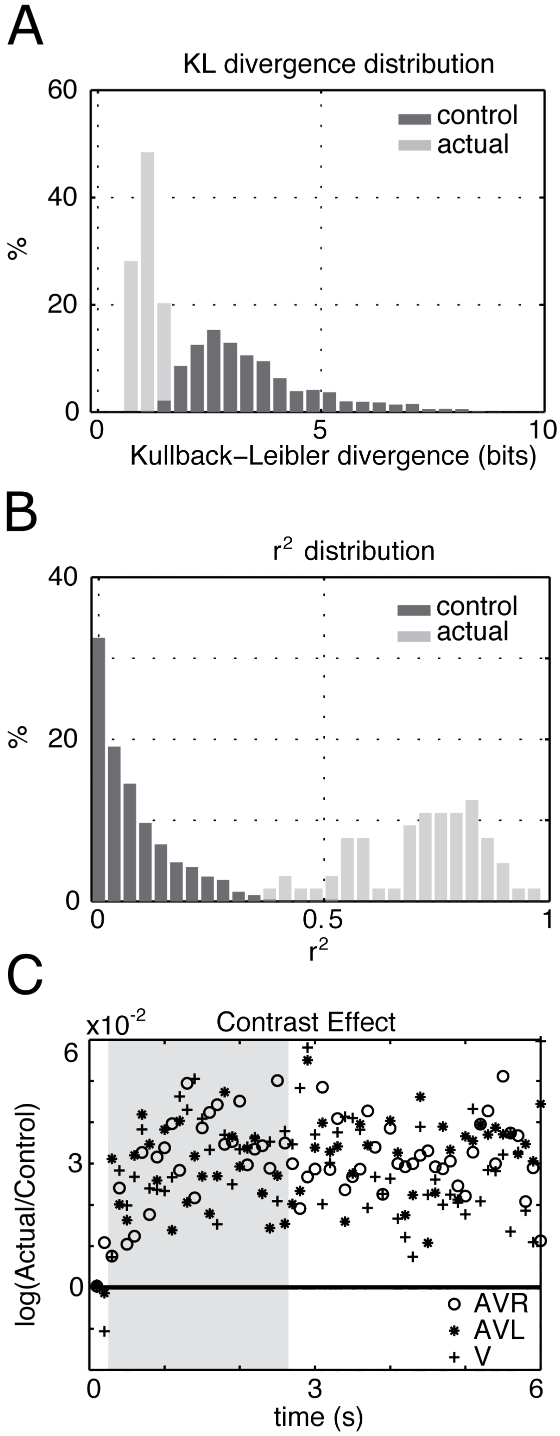

Figure 5.6: **Auditory and Visual Information are Integrated.** The distributions of KL divergence **(A)** and r^2 **(B)** values for control (textitdark gray) and actual (*light gray*) conditions are shown. The actual distributions are obtained by comparing 64 pairs of multimodal (AVR and AVL) and unimodal (V) fixation pdfs. Control distributions are created by computing the same statistics on randomly-paired non-matched multimodal and unimodal pdfs (n = 3200). The measurements are directly obtained from pdfs shown in Fig. 5.5. **(C)** The logarithm of the ratios of actual and control luminance contrast values are presented as a function of time. Almost all values lie above the identity line. This effect is stable over the time of presentation and for different conditions. The *gray shaded* area shows the temporal region of interest.

nance contrast. Moreover, the effect of contrast is constant over the entire presentation, with no systematic difference during the temporal interval of interest (Fig. 5.6C, gray area). Furthermore, the three conditions do not differ significantly in the size of effect. This holds for temporal bin sizes ranging from 100 to 1000 ms (data not shown). Therefore, we can conclude that the additional presence of a lateralized auditory stimulus does not reduce the correlation between the subjects' fixation points and luminance contrast.

5.4.4 The Integration Is Linear

To quantitatively characterize the integration of unimodal saliencies, we perform a multiple regression analysis. The experiments involved two unimodal auditory conditions (AL and AR), and two corresponding multimodal conditions (AVL and AVR). In order to simplify subsequent discussion, we will use a more general notation of A and AV for unimodal auditory and multimodal conditions respectively. We model the multimodal (AV) distributions of fixation points by means of a linear combination of unimodal (A and V) distributions and the normalized (to unit integral) product of unimodal distributions (A ×V) as indicated in equation (5.6). Here we assume that the effect of a multiplicative integration of unimodal saliencies is well approximated by a simple multiplication of their probability distributions. The fits were computed separately for left and right conditions, and the results were pooled for visualization purposes. The distribution of 64 coefficients for each of the regressors is shown in Fig. 5.7. On average, over all images, the contribution of unimodal visual saliency is largest, with a mean at 0.75 ± 0.15 (mean ±SD; Fig. 5.7, circles). The contribution of unimodal auditory saliency is smaller (0.16 ± 0.12). The coefficient of the cross-product interaction term is, however, slightly negative with mean 0.05 ± 0.10. We repeated the same analysis for a subset of subjects (n=14, 40 %) for whom the lateralized auditory stimulus had the strongest effect on fixations in terms of gravity center shift in the unimodal auditory conditions. In these subjects, the contribution of auditory coefficients was increased (0.32 ± 0.17) at the expense of the visual ones (0.53 ± 20), without any apparent effect on the interaction term (-0.06 ± 0.11, t-test, $p = 0.59$) (Fig. 5.7, crosses). In both cases, the intercept was very close but still significantly different from zero. These results suggest that biggest contributions to the multimodal pdfs originate from the linear combinations of the unimodal pdfs.

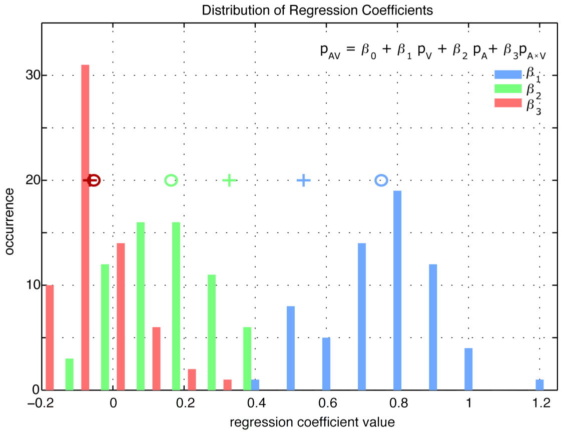

Figure 5.7: **Prominent Contribution of Unimodal Saliencies.** This figure shows the distributions of the parameters of equation Fig. 5.6 (inset), computed using multiple regression analysis. Image pdfs used for the regression analysis were computed either by using all subjects (n = 35) or by selecting a subset of subjects (n = 14) for whom the lateralized auditory stimulus had the strongest effect on fixations, quantified in terms of gravity center shift. The best fits between image pdfs from conditions AV and V, A and A ×V were calculated for each of the 32 images, in left and right conditions, yielding 64 fits for each parameter. The distributions of the coefficients resulting from the regression analysis using all subjects are shown. *Circles* denote the mean of each distribution of each coefficient. All means are significantly different from zero (t-test, p $<10^{-3}$). The average value of the visual coefficients $\beta_1(0.75 \pm 0.15; \pm\text{SD})$ is greater than the average of the auditory coefficients $\beta_2(0.16 \pm 0.12)$, and these unimodal coefficient averages are both greater than that of the multimodal coefficients $\beta_3(-0.05 \pm 0.1)$. Repeating the same analysis using the subset of auditorily-driven subjects results in higher average auditory coefficients (0.32 ± 0.17) and lower visual coefficients (0.53 ± 0.20), while the interaction term does not change significantly ($-0.06 \pm 0.11, t{-}test\text{p} = 0.59$). *Crosses* indicate the means of each and every distribution for this subset of subjects.

In a subsequent analysis, we carried out the regression analysis using a different combination of dependent variables and evaluated how the introduction of an additional dependent variable increased the explained variance. Using only unimodal visual pdfs as one regressor, we obtained r^2 values having a median value of 0.72 over all images —as expected from the previous section. Additionally including unimodal auditory pdfs increased the median r^2 only slightly, by 3%. Repeating this analysis using only the subset of subjects showing the strongest auditory lateralization effect, we obtained a median value of 0.36 with the sole visual regressor. The subsequent introduction of the unimodal auditory pdfs as second dependent variable increased the goodness of fit by 21% over all images. Further including the cross- interaction term as the third dependent variable, the goodness of fit increased slightly, by 5%. Therefore we can argue that mechanism linearly combining the unimodal saliencies can well account for the observed behavior in the multimodal conditions.

As a model-free approach, we compute integration plots using saliencies obtained from different conditions. It should be noted that no assumptions are made regarding the calculation of the saliency; that is, these saliency values are empirically determined by the gaze locations of many subjects. Integration plots are constructed by plotting the saliency of a given spatial location in the multimodal pdfs as a function of unimodal saliencies of the same spatial location. The specific distribution within this three dimensional space describes the integration process. In Fig. 5.8 A, B and C, the height of each surface depicts the corresponding salience of the same location during the multimodal condition as a function of the saliency of the same location in unimodal conditions. The three hypotheses about the integration of auditory and visual information make different predictions (see introduction): Early interaction leads to a facilitatory effect and an expansive non-linearity (Fig. 5.8). The landscape predicted in the case of linear integration is planar and shown in Fig. 5.8. Late combination gives rise to a compressive non-linearity (Fig. 5.8C). Applying this approach to our complete dataset leads to a highly non-uniform distribution of observed unimodal saliencies, when considered in terms of their frequency of occurrence. In Fig. 5.8D, the count matrix of joint occurrences of unimodal saliencies is presented; the surface is very sparsely filled at regions with high values of salience and practically no point is present at regions where both unimodal saliencies are high (top right region of Fig. 5.8D). Statistically reliable statements about the data within these regions are thus not possible. Within this space, we defined a region of interest depicted as a rectangle in Fig. 5.8D; this portion of the space contains 85% of the total number of samples. Inside this region, we bin the data using a 10 ×10 grid and calculate the expected value (Fig. 5.8) and an error estimate (variance) for each bin.

The relationship between these unimodal and multimodal saliencies is further analyzed using a weighted regression analysis with unimodal saliencies as dependent variables. This yielded 0.54 ± 0.04 ($\pm$ 95 %CI) and 0.59 ± 0.09 for the linear contribution of visual and auditory saliency respectively. Both coefficients were highly significant (t-test, p $<10^{-6}$)except for the intercept coefficient (t-test, $p = 0.23$). r^2 is equal to 0.89 suggesting a good fit. We repeated the same analysis with the interaction term included after normalizing each regressor with its geometric mean in order to have the same exponent range, thus permitting an evaluation of the contribution of different regressors to the multimodal saliency. This yielded 0.57 ± 0.08, 0.29 ± 0.08 and 0.029 ± 0.05 for

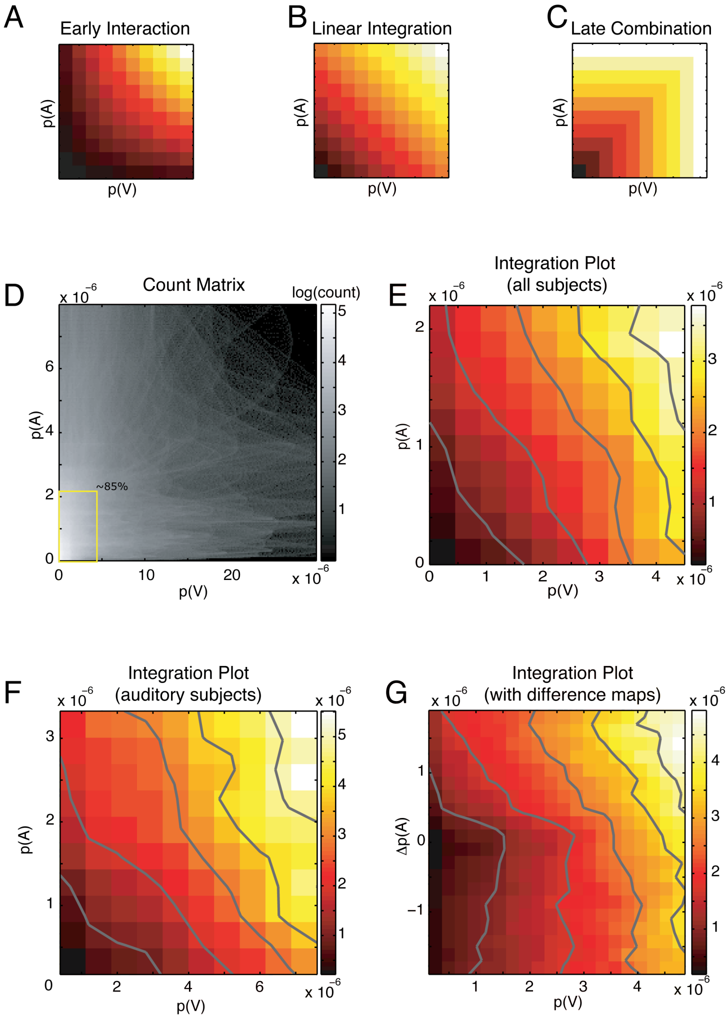

Figure 5.8: **The Integration is Linear.** **(A, B, C)** Three hypothetical frameworks for integration are presented schematically as integration plots. The X and Y axes represent the unimodal saliencies associated with a given location on the visual field. The saliency of the same location in the multimodal condition is color-coded, i.e. each pixel represents the saliency of a given point in multimodal condition (p(AV)) as a function of the saliency of the same point in unimodal conditions (p(A), p(V)). The specific distribution of the points generating this landscape unravels the integration process. Please note that the inherent topology of underlying images is no longer contained in integration plots. The three integration schemes mentioned in the text (see Introduction) predict different probability landscapes. If the unimodal probabilities were interacting this would generate a landscape with an expansive non-linearity **(A)**; however if the multimodal saliencies were combined linearly, the resulting landscape is expected to be planar **(B)**. The absence of an integration in a scenario where the maximum of the unimodal saliencies determines the multimodal saliency results in a compressive non-linearity **(C)**. **(D)** Joint count matrix obtained by using all subjects. All pairs of unimodal saliencies are plotted against each other. Grayscale level logarithmically codes for the number of occurrences inside each bin. The marked rectangular region contains 85% of all points. **(E)** Integration plot calculated for points lying in the rectangular region of (D) using a 10 ×10 binning. The color codes for the saliency in the multimodal condition as in (A-C). Image pdfs used to compute this plot are obtained by using all subjects. **(F)** Same as in (E), however only subjects with the strongest auditory response are used. **(G)** Integration plot, calculated using difference maps (see Results for details).

visual, auditory and interaction terms respectively. The linear contributions of unimodal saliencies were highly significant whereas the intercept and the interaction terms were not statistically different to zero.

Using such large bins increases the statistical power within each bin at the expense of detailed structure. We therefore conducted the same analysis using up to 50 bins covering the same region of interest. The r^2 of the fitted data at this resolution was 0.87, ensuring that the fit was still reasonably good. The values of coefficients were practically the same, and the only noticeable change during the incremental increase of the resolution was that the interaction term reached the significance level ($p > 0.05$) at the resolution of 20 ×20, thus demonstrating a slight facilitatory effect. These results support the conclusion that linear integration is the dominating factor in crossmodal integration during overt attention, with an additional small facilitatory component.

The above analysis is influenced by a particular property of the auditory saliency maps. Many fixation densities in the AL and AR conditions are located at the center of the screen (see Fig. 5.4C). We tried to avoid this problem in two different ways. In the first method, we performed the same analysis on the subset of subjects mentioned earlier who were most influenced by the lateralized auditory stimulus, thus minimizing the central bias. Restricting the analysis allowed us to define a new region of interest, which included 90% of the total data points (Fig. 5.8E) and discarded only those points that were very sparsely distributed in high saliency regions. r^2 values varied within the range of 0.81 and 0.9, decreasing with higher binning resolutions. As above, increasing the number of bins revealed a slight but significant facilitatory effect. Within this subset of subjects, the contribution of auditory saliency (0.36 ± 0.08) was again shown to increase at the expense of the visual contribution (0.50 ± 0.08). Removing the interaction term from the regression analysis caused a maximum drop of only 2.5% in the goodness of fit for all tested bin resolutions within this subset of subjects.

In the second method used to remove the central bias artifact of unimodal auditory pdfs, we took the differences between the left and right auditory conditions, i.e. subtracted the two empirically determined auditory saliency maps to yield difference maps. In each case, the saliency map for the condition where the sound was presented contralaterally was subtracted from the saliency map of the congruent side (i.e. AL –AR for auditory stimulus presented on the left, AR –AL for auditory stimulation from the right). These newly generated maps are well-behaved and allow the analysis of a larger region of saliency space (90% of total samples). The above analysis was repeated using the difference maps, and is shown in Fig. 5.8G. It should be noted that the positive values on the y-axis are the data points originating from the region of the screen from which the sound emanates, during the temporal interval of interest. We performed separate regression analyses for these two halves of the resulting interaction map. In the lower part ($p(A) > 0$), the best predictor was the visual saliencies, as can be seen from the contour lines. In the upper part ($p(A) < 0$), a linear additive model incorporating auditory and visual saliencies well approximates the surface. The results derived in an analysis of the effects of different bin sizes were comparable to the above results; that is, a model combining linearly unimodal saliencies along with a slight facilitatory component was sufficient to explain a major extent of the observed data.

We repeated the last analysis with image pdfs obtained with varying degrees of smoothing. Decreasing the width of the convolution kernel systematically reduced the explained variance of the fits on integration plots built within probability ranges containing comparable amount of points. In a large interval of tested parameters (0.4° –0.8°), the main result was conserved i.e. the saliency surface was, to a large extent, captured by a linear combination of unimodal saliencies, with a slight multiplicative effect also evident.

5.5 Discussion

In this study, we investigated the nature of the multimodal integration during overt attention under natural conditions. We first showed that humans do orient their overt attention towards the part of the scene where the sound originates. This effect lasted for the entire period of the presentation of the stimuli, but had a stronger bias during the first half of presentation. More interestingly, this shift was far from a simple orientation behavior —overt behavior during multimodal stimuli was found to be dependent on the saliency of both visual and auditory unimodal stimuli. Although subjects' fixation points were biased towards the localized auditory stimuli, this bias was found to be dependent on visual information. Our analysis suggests that a predominantly linear combination of unimodal saliencies accounts for the crossmodal integration process.

We quantified the saliency associated with a given image region by analysis of the measured overt eye movements of a large amount of subjects. Subjects' behavior was similar within the temporal interval where the effect of the lateralized sound was strongest. However we do not know whether this was the result of a search strategy shared between subjects or whether it originates purely from the bottom-up content present in the image. The results presented here do not depend on the precise determinants of the saliency associated to different image regions. Similarly, we evaluated the saliency of different parts of the visual field associated with the lateralized auditory stimulation. In many subjects this resulted in a shift of fixation toward the side of the sound, in accord with previous studies showing that sound source location is an important parameter.

Prior studies (Corneil & Munoz 1996, Corneil et al. 2002, Arndt & Colonius 2003) have shown that congruent multimodal stimulation during tasks where subjects were required to move their gaze to targets as fast as possible results in faster saccadic reaction times together with an increase in the accuracy of saccades. Here we are extending these results to more operationally relevant conditions by using natural stimuli under free-viewing conditions where the subjects are not constrained in their behavior. Moreover, we are formally describing the crossmodal behavior in terms of unimodal behavior.

Concerning the temporal dynamics of the integration process, we found that the localized sound stimuli attracts subjects' attention more strongly during the first half of presentation, corresponding to an interval of approximately 2.5 seconds. Although it is observed that the lateralization of fixation density continues throughout the whole presentation time (Fig. 5.4), the effects are much weaker and do not reach the significance level. This effect can be understood as a consequence of inhibition of return, the subject losing interest in that side of the image, or alternatively due to an increasing efficiency of top-down signals over time, resulting in a superior efficiency of the sensory signals to attract attention during early periods of exposure only.

One interesting point is to know whether the present results —a linear integration of auditory and visual saliencies —generalize to situations with a combination of complex visual and complex auditory scenes. In the proposed computational scheme the origin of visual and auditory saliency maps is not constrained, but measured experimentally. The spatial structure of the auditory salience map is more complex, but presumably does not much the spatial acuity of the visual system. As a consequence, in the case several auditory stimuli would contribute no fundamental property in the integration process needs to be changed and we expect the same integration scheme to hold.

In our study, majority of the natural images we have presented to the subjects were devoid of human artifacts. It could be argued that our auditory stimuli were semantically more congruent with natural scenes where there was no human artifact visible and therefore the crossmodal integration would be stronger. Although some arbitrary decisions has to be taken, we separated our visual stimuli into two classes depending on whether human artefacts were present or not, and conducted the regression analysis with these two sets separately. We have not found stronger integration in the case of natural images without human artefacts compared to the case where human artefacts were visible.

How do these results fit with current neurophysiological knowledge? One of the most studied structures in the context of cross-modal integration is the superior colliculus, a deep brain structure (Meredith and Stein, 1986; Stein et al. 2004). It has long been known that superior colliculus contains neurons that receive inputs from different modalities. Neurons fire more strongly with simultaneous congruent spatial stimulation in different modalities, compared to unimodal firing rates. A recent report (Stanford et al., 2005) which attempted to quantify this integration process occurring in the superior colliculus, has pointed out that a great deal of the integration can be described by the linear summation of the unimodal channels, thereby providing supporting evidence that it is possible for linear integration to be implemented in the brain. At the cortical level, we are far from obtaining a final clear-cut consensus on how saliency is computed and integrated. In order for a cortical area to fulfill the requirements of a saliency map, the activity of neurons must predict the next location of attentional allocation. A number of such cortical areas have been proposed. Primary visual cortex, the largest of all topographically organized visual areas, may contain a saliency map (Li, 2002). Simulations inspired by the local connectivity of V1 generate results compatible with human psychophysical data, thus linking the activity of neurons in early visual areas to the computation of salience. By recording single unit activity in monkey cortex during the exploration of natural visual stimuli,Mazer and Gallant (2003) found that the activity of neurons in V4 predicted whether a saccade would be made to their receptive fields. Based on these findings, they argue that V4, a higher level area located in the ventral visual stream, contains a topographic map of visual saliency. It is likely that these considerations may be generalized to other areas located in the ventral stream, but presumably also to cortical areas responsible for auditory processing. In addition, areas in the dorsal visual pathway and the frontal lobe —lateral intraparietal (LIP) area and frontal eye field (FEF) respectively —have been associated with saliency. The activity of FEF neurons can be effectively modulated by the intrinsic saliency of the stimuli and further modulated by the current requirements of the task (Thompson and Bishot, 2004; Thompson et al., 2005). In the dorsal pathway, Bisley and Goldberg (2006) propose that LIP displays the crucial properties of a saliency map. Since saliency related activity in the brain seems to be widely distributed over many areas these areas could in theory be in the position to independently compete for the control of overt attention. However, our results support the existence of a joint functional saliency map, in which the information from different modalities converges before the non-linearities involved in the process of fixation point selection are applied.

It should be noted, however, that our results can not unravel the neuronal mechanisms underlying integration, as the exact cellular computations could in principle be carried out by operations other than linear summation of local variables. This depends on how saliency is represented —for example, if saliency were represented logarithmically, the linear summation would create a multiplicative effect. What we have shown is that at the behavioral level the information converges before motor decisions are taken and that this integration is mostly linear. We thus provide boundary constraints on the computations involved in the control of overt attention.

Renninger et al. (2005) use information theoretical tools to provide a new framework for the investigation of human overt attention. According to this hypothesis, the information gain is causally related to the selection of fixation points, i.e. we look where we gain the most information. Considered within this framework, it is tempting to speculate that a linear integration scheme of unimodal saliencies is compatible with the optimal information gain, in the sense that the linear integration of information gains originating from different modalities provides the optimal combination strategy, as the information gain is the sum of the information quantities that each modality provides.

The integration of multiple sources of information is also a central issue in models of attention operating in unimodal conditions. As already mentioned, modality-specific information is separated into different feature channels, and the subsequent integration of these different sources is usually subject to arbitrary decisions on the part of the modeler due to the lack of biologically relevant data arising from natural conditions. Whether unimodal feature channels are also linearly integrated is a testable hypothesis and needs further experimental research.

One problem we encountered was the centralized fixation density present in the fixation probability distributions of unimodal auditory conditions. Although subjects were effectively oriented by the lateralized sound source, most of their fixations were concentrated at the center of the monitor. We avoided this problem in our analysis by taking the difference of the probability distributions obtained in unimodal auditory left and right conditions, and also by constraining analysis to the subset of subjects whose behavior was most influenced by auditory stimulation. However, we believe that this problem may be alleviated by using multiple sounds simulated to originate from different parts of the image.

It is common for complex systems, composed of non-linear units, to function in a linear way. Neurons are the basic functional constituents of nervous systems, and may express highly non-linear behaviors; for example, the Hodgkin-Huxley equations describing the relation between membrane potential and ionic currents are highly non- linear. Furthermore, the excitatory and inhibitory recurrent connections within and between cortical areas allow for complex non-linear interactions. However, irrespective of these underlying non-linear aspects, many neuronal functions are still well described by linear models. Neurons in early sensory cortices (Schnupp et al. 2001) such as simple cells (Carandini et al. 1997), for example, are well approximated when considered as linear filters operating on input signals. A recent study involving micro-stimulation in motor cortex showed that signals for movement direction and muscle activation also combine linearly. (Ethier et al. 2006). We have shown that the cross-modal integration during overt attention process is best described as a linear integration of sensory information, possibly originating from different brain areas. In doing so, we have provided an important constraint for any model of cross-modal interaction. This raises an important challenge for any biologically plausible model of human overt attention operating in environments with multiple source of information.

6

General Discussion

The main axis of this thesis focused on approximating real-world operational conditions within the laboratory environment by using complex and rich natural stimuli. To this end, a considerable amount of energy was devoted to gathering these stimuli, which were subsequently used in a variety of experimental and theoretical settings. Together with many recent reports published during the last few years (see Felsen & Dan (2005) for a review), this thesis provided supporting evidence that the usage of natural stimuli for the study of central nervous system is both possible and productive in empirical and theoretical terms, thereby speaking against a large number of neuroscientists, who are rather reluctant for the entrance of natural stimuli into experimental fields due to their complex and uncontrolled nature (Rust & Movshon 2005).

The central nervous system of mammals contains many more synapses than the number of genes that exist in their genome. For example, in humans the number of genes is estimated to be in the order of 50,000 and no estimation returns a number larger than 100,000 (Venter et al. 2001). These estimations are 10 orders of magnitude smaller than the number of synapses in the human brain (Herculano-Houzel 2009). Assuming this ratio as a valid rule of thumb for other species, a genetic determination of the fine connectivity patterns between specific neurons seems to be implausible. Therefore, a generic versatile learning rule which does not depend on explicit genetic definitions, is extremely useful for achieving a given level of selectivity (e.g. edge orientation) with respect to the external world. Moreover, absence of the necessity of a hard-wiring strategy via genetic information enables considerable plasticity with respect to relevant features that are specifically needed by different animals.

Learning stable features from the input signals is a good candidate for achieving this goal. We reported that the binocular disparity coding, observed in the primary visual area of monkeys and cats can be understood in terms of extracting stable features from input signals, given that these are rich in content and represent the natural input. As this learning scheme involves the temporal dimension, the result of slow feature analysis depends crucially on the properties of the body and motor repertoire of the animal in question and enables neurons to be selective for different features which are relevant for a given animal. For example, the neurons of animals with fast versus slow locomotion can extract features suitable for the range of behaviors that their bodies allow without any explicit definitions in their genomes being necessary.

In addition to low-level feature selectivity that neurons in the early sensory cortices express, the operations carried out across the visual ventral pathway involve considerably complex computations. For example, neurons located in the temporal cortical areas of monkeys have RFs that are selective for complex patterns representing specific objects. Moreover, this selectivity is to a large extent invariant to the viewing angle of objects and their position in the visual field (Gross et al. 1972, Rolls & Deco 2006). Even higher in the hierarchy across the visual ventral pathway, the hippocampus occupies the final point of neuronal convergence of a large-number of cortically interconnected areas. Some of the hippocampal neurons called place cells (O'Keefe & Dostrovsky 1971) fire when an animal is at a given specific location in the space irrespective of the precise viewing angle of the animal in that portion of the space. This means that visual input is processed in such a way that these neurons extract the view-angle-invariant information from the input signals. In simulation studies such non-trivial RF properties were obtained using the same temporal stability based learning rule. For example, concerning the temporal cortical neurons, Einhäuser et al. (2005) showed that the translation invariant object recognition can be achieved by extracting stable features from a series of object pictures taken from different angle of view. Furthermore, Wyss et al. (2006), Franzius et al. (2007) provided evidence that extracting stable features in an hierarchical manner from the visual signals recorded by a robot exploring either a real-world (Wyss et al. 2006) or a virtual environment (Franzius et al. 2007), place-selective neurons emerge at the top layers. It is therefore likely that the functioning of the ventral visual pathway, however complex the computations carried by these neurons are, may indeed be based on a simple learning rule iterated over and over across cortical areas.

Cortical neurons also exhibit many different other specializations and there is no doubt that the complete operation of the cortical network constituting the visual areas cannot be totally captured by a simple formal expression. For example, the hierarchical organization of visual areas reflects only one side of the facts, because it is well known that the bidirectionality in the cortico-cortical connections is rather a general rule of organization (Felleman & Essen 1991). These connections are also distributed across different modalities, effectively multiplying the integrative power of neurons. Moreover today, research on crossmodal integration provides experimental evidence suggesting that different sensory pathways classically thought to be segregated —at least in the early sensory areas —may actually overlap considerably, resulting in the integration of different sensory signals very early in the hierarchy (Dehay et al. 1988, Driver & Noesselt 2008, Macaluso et al. 2004). Adding to the complexity, cortical areas also contain a dense web of intra-areal connections, which are highly organized (Rockland & Lund 1982). According to our current knowledge, as early as in the primary visual area, the organization of these local connections are thought to reflect the general statistical properties of the input signals founding the basis of local integrative processes (Betsch et al. 2004). There is no doubt that more experimental research is needed in order to understand better the cortical function.

Using the physiological recording technique of VSDI (Grinvald & Hildesheim 2004), we aimed to understand the processing of natural input in the visual area. We were able to analyze the activity of a relatively large cortical area with reasonably good spatial and temporal resolution. Even though the spatial resolution of this technique is not strong enough to resolve single neurons, it however provides the possibility of recording a large number of neuronal masses directly and simultaneously. Importantly, it captures the global cortical state spanning many columns and pinwheels, and therefore it is much less susceptible with respect to any bias that the experimenter may introduce inadvertently during electrophysiological recordings. For example, whereas it is simply not possible to record electrophysiologically from a silent neuron, VSDI would not be blind to their presence as it faithfully captures the big picture.

We found major differences in the cortical activity levels evoked by natural versus simple laboratory stimuli. The physiological measurements indicated that during the processing of commonly used simple drifting edge stimuli, the visual cortex is driven toward a state characterized by an excess of excitation, leading the system into a strong hyperexcitatory state. The activity levels in response to natural stimuli were much more moderate in comparison. We interpreted these results as being the consequence of an evolutionary adaptational process which leads to the efficient processing of naturally occurring sensory inputs in terms of energy consumption. Within this framework, the fact that the processing of natural stimuli does not require high levels of net excitation can be conceived as the result of specialization of the cortical circuits.

Beside these observed differences in responses to artificial and natural stimuli, it is also important to emphasize that these discrepancies may create difficulties for interpreting the conclusions derived from studies where simple, artificial stimuli are used. Let's take as an example, the emergence of orientation selectivity in the primary visual cortex. It is by far the most investigated case study in cortical computation since the early work of Hubel and Wiesel (Priebe & Ferster 2008). However, many of these studies are based on the results obtained by laboratory stimuli. Taken together with the results presented here, it is reasonable to doubt the generalizability of these results. For example, the orientation selectivity measured under artificial conditions, may well be an underestimated approximation of the real selectivity of neurons. The excess of excitation in response to grating stimuli (or/and the excess of inhibition in response to natural movies) could in principle lead to such differences in the tuning of neurons. These considerations should lead us to be more prudent for the interpretations of results.

In the experimental paradigms which are currently used, the final cortical response is obtained after averaging a large number of repeated trials, thus averaging out the noise component. By analyzing the data in a trial by trial basis, (Arieli et al. 1996) has shown that a considerable part of the cortical variability in response to stimulation is due to the spontaneous activity levels, rather than the external stimulation. This means that the complexity of the signals can only be underestimated with the experimental paradigms we are currently using. We showed that even following averaging a large number of trials, complex spatio-temporal activity dynamics constitute a fundamental characteristic of the responses of early sensory cortex to input signals, be it complex or simple in nature. Propagating waves were consistently observed during both grating and natural conditions. This fact raises the question of how a system characterized by complex dynamics can self-coordinate in order to integrate different sources of information. We found that these complex dynamics representing the sensory input were, besides being locked to the dynamics of the external stimuli, also very sensitive to internal signals conveying contextual intra- or extramodal information. The presence of contextual visual information was found to support the activity levels of distant neuronal populations. In the same vein, an additional auditory signal was found to increase the strength of the motion locking of the distant visual cortical areas. Therefore it is tempting to state that the cortical machinery is extremely prone to integrate different sources of information. Given this strong integrative power, it is however remarkably robust in conserving its stability under normal operating conditions.

As the presence of integrative phenomenon at the neuronal level does not determine unambiguously the nature of the motor output at the behavioral level, the effectiveness of the integrative processes can be best studied in the behavioral level. Interestingly, the results of my psychophysical experiments paralleled the physiological results. Instead of having different sensory modalities in competition with each other for the control of the behavior, we reported that different signals were integrated before the production of the motor decision. Remarkably, the integration scheme was well predicted by a linear combination of different sources of information. It is surprising that the presumably complex dynamics underlying the functioning of neuronal populations, results in interactions which are predicted by simple linear models on the behavioral level.

This thesis coincided with a transition period in systems neuroscience, a transition period consisting of the introduction of more and more elaborate, complex and clever experimental conditions within the laboratory settings. In my opinion the usage of complex natural stimuli for the investigation of sensory systems represents only the starting point of this global trend. Yet many exciting findings have already been brought to light concerning the adaptational mechanisms used by cortical neurons in sensory cortices. As the central nervous system of mammals is evolved for maintaining of the survival in face of extremely uncertain environmental conditions, these findings may not be utterly surprising from an evolutionary perspective. For example, the presence of a prey, once detected by sensory neurons, must be reliably communicated to other regions of the brain, and in order to avoid unnecessary energy consumptions and to guaranty the survival, the probability of a prey or a predator at a given location needs to be evaluated pertinently based on very limited number of trials (Glimcher 2003). Therefore very robust neuronal mechanisms need to be at play to evaluate the risks of actions and their outcomes. However, such mechanisms can not be uncovered when the complexity of the our experimental setups do not match to the complexity of our subjects. Therefore conducting better and richer experiments in the future will be our chance to learn new insights about the way the brain functions.

7

Appendix

7.1 Description of Optical Imaging Data & Generic Processing Steps: From Photon Counts to Neuronal Activity

VSDI is a recording technique for directly measuring the activity of large amounts of neuronal masses with a temporal resolution in the order of milliseconds. It combines relatively good temporal resolution with a satisfying spatial resolution. However as it is always the case, each recording, be it physiological or physical, has its own peculiarities due to the interaction of the recording system with the subject of the investigation. VSDI is not exempt of problems and needs careful preprocessing before any analysis. In the following, description of the data and conventional preprocessing techniques will be explained in detail.

7.1.1 Some Terminology

Introducing the following definitions is necessary in order to understand the terminology used in this Section. Due to the manner in which experiments are done, we have to distinguish between *trials*, *trialblock* and *block*. A *trial* is the shortest data segment that is recorded. A *trial block* is a set of trials where all possible conditions are recorded in a randomized fashion. In the experiments, we have 17 conditions. The data coming from the recordings of 17 conditions compose a *trialblock*. This is the unit of data in the hard disk because a given trialblock is stored as a unique file. The repetitive recording generating many trialblocks creates a *block*. During a block, nothing in the recording hardware (such as gain control, aperture etc.) is modified and as a consequence a block is an ensemble of trialblocks where many trials are recorded in presumably constant conditions. However this situation is interrupted due to some necessity which forces the experimenter to change some parameters in the camera system. As a consequence a new block starts after parameter adjustments.

In summary, a trialblock refers to a set of trials where all the stimulus conditions are shown only once. A block refers to a group of trialblocks where the hardware parameters are kept constant. A new block starts as soon as a parameters are modified.

7.1.2 Properties of Raw Data

Each trialblock can be conceptualized as a time-series organized in a 3 dimensional matrix as (space,time,condition). The length of the time dimension depends on the specific sampling rate used during a given experiment. The presentation duration is always 2 seconds. An interval of time corresponding one tenth of the whole stimulus presentation duration corresponds to the pre-stimulus period.

An example frame of the raw data, directly output from the camera, is shown in Fig. 7.1A (left panel). The color coded numbers are proportional to the number of photons received by the photosensitive elements of the camera facing to the cortical tissue depicted in the right panel. The uneven distribution of the luminance values is clearly visible as an inverted basin-like structure. Three artefactual lines in the raw data are notable. These lines are due to the known artefacts caused by the camera system used. Importantly we see that the cortical "picture" at this rather raw level is mainly dominated by the curvature of the cortex which causes a strong spatial inhomogeneity on the number of photons recorded. These artefacts and the spatial inhomogeneities will be washed out by the subsequent preprocessing steps discussed below.

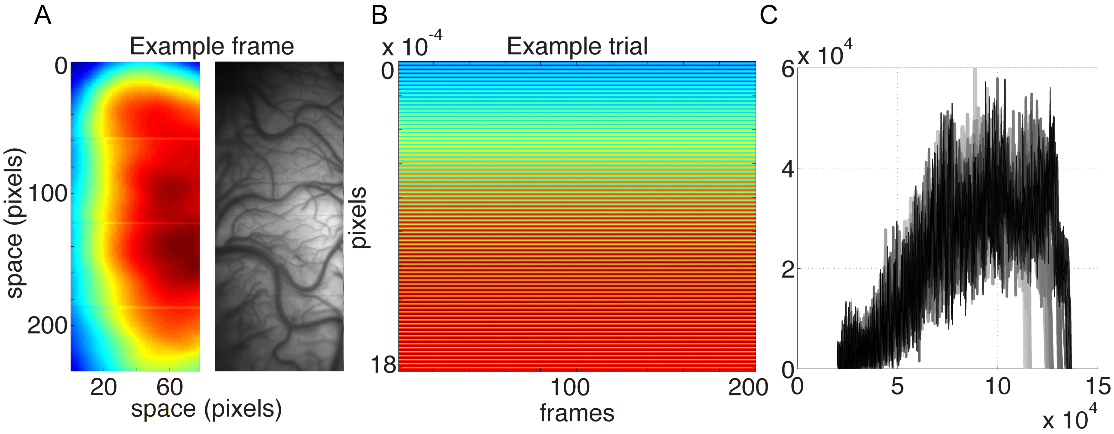

Figure 7.1: **Raw Data Directly Output from the Camera.** **(A)** One example of recorded frame is shown on the right panel. The color coded values are proportional to the number of photons received. On the right panel, the corresponding cortical tissue is shown. **(B)** The temporal evolution of the raw data. Please note that the change across pixels is much bigger than the change occurring across time. At this initial stage no trace of evoked activity is visible. **(C)** Histogram of the luminance levels, each line represents a trialblock. The color and thickness of the lines changes according to how late a given recording is done. *Thick* and *bright* lines indicate early recordings and *black* and *thin* lines indicates late (toward the end of the experiment) recordings. To better see the evaluation of the slow drift over the recording period refer to the next figure.

In Fig. 7.1B, temporal evolution of each pixel's luminance is depicted (in units of recorded frames each lasting 5 ms). The data presented in Fig. 7.1A (left panel) occupies one of the columns in this panel. Importantly the deviations from the baseline due to the stimulus onset (which occurs at 20th frame) are not visible. The reason for this is

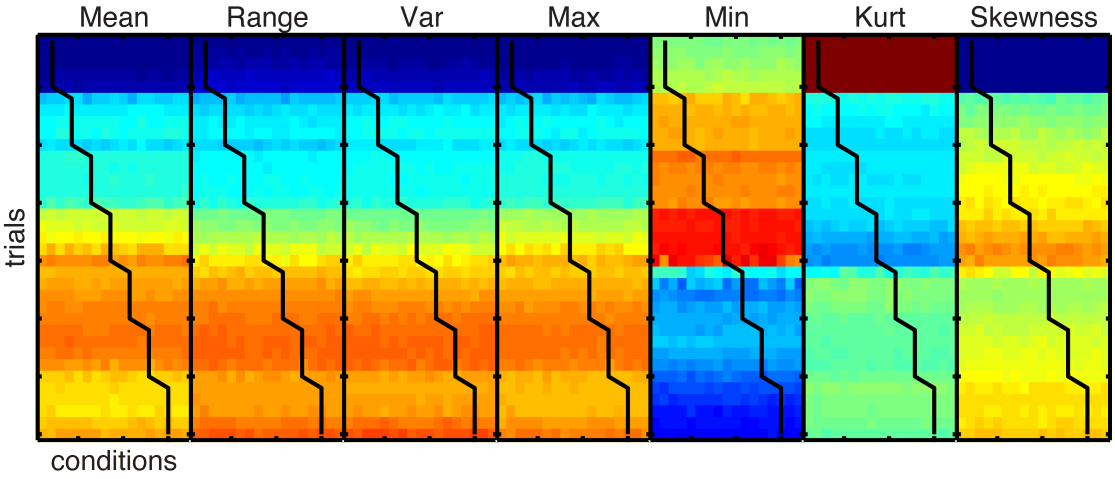

Figure 7.2: **Fluctuations in the Basic Statistics Extracted from the Raw Data of a Given Experiment.** On the left-most panel, the average value of the recorded raw signal is shown for all conditions (x-axis) and trials (y-axis). Different panels represent different statistics extracted. The staircase black lines depict the block identity changes, each time a step occurs a new recording block starts.

that the deviations between the DC components (average values) of different pixels are comparatively much bigger than the effect of the stimulus onset.

In Fig. 7.1C, we see the distributions of these raw luminance values. Each line corresponds to a given trialblock. All the pixels which makes up a trialblock (therefore pooled over conditions, space and time) were used to create each single histogram. Here the brightness of the curves represents different trials. Brighter and thicker lines represent trials which are recorded earlier. The fact that late recordings within this experiment have more illumination is caused by the experimenter's adjustments counter-balancing the photo-oxidative degradation of dye signals.

Different statistics extracted from the dataset recorded during an experiment are shown in Fig. 7.2. In the y-axis, we see different trials recorded along the experiment ($\approx$35 trials). In the large-scale x-axis different statistics are represented: *mean, range, variance, maximum, minimum, kurtosis, skewness.* These statistics are extracted for each condition separately by pooling all the data points together. Within each subplot, across the x-axis different stimulus conditions are shown (total of 17). The black thin line symbolically shows the starts and stops of different blocks. The intervention of the experimenter are clearly visible in the abrupt changes of these statistics. It is expected that an ideal preprocessing method cancels out these fluctuations which are dependent on experimenter's interventions.

7.1.3 Frame-zero Correction by Division

Frame-zero correction is realized in order to remove differences in the DC values of different pixels. This results in the correction of spatial inhomogeneities. In order to do so, we artificially equalize the luminance values recorded during the prestimulus period. During this process, for each trial data, the average luminance during the pre-stimulus time is computed for each pixel and condition separately. This generates one scalar value for each pixel and condition. Following, all recorded samples belonging to a pixel and condition are normalized by this scalar. This forces the temporal average during the pre-stimulus period to be equal to 1 for each of the pixels. This is called *frame-zero* correction.

7.1.4 Why Division?

Even though it is consistent with the literature, it is not clear why division rather then a subtraction need to be the preferred method. In a scenario where the dynamic range of a given pixel is proportional to its prestimulus DC value, the divisive normalization would be preferable. However subtraction would be preferred if the gain of the activation is constant and therefore independent of the DC component.

This can be verified easily. Fig. 7.3 shows for a given trial and for each pixel, the relationship between the DC component and its dynamic range. Here dynamic range of a given pixel is quantified as the standard deviation of its activity computed across the temporal dimension. This was done for two different stimulation conditions (grating condition (blue) and blank condition (red)). It is clearly visible that the modulation depth is to a good extent linearly related to the DC component of the pixels. This justifies the necessity for a divisive normalization procedure.

The effect of the divisive frame-zero correction on the raw data is presented in Fig. 7.4A (compare to Fig. 7.1B to the raw data). The frame-zero division has the effect of making the distributions of luminance values more Gaussian-like (compare histograms in Fig. 7.4C and 7.1C). In these histograms all of the distributions are centered around a value close to 1. Although some drifts over the course of the experiment (a given gray level codes for a given trialblock) are present, there seems to be no general rule governing how do these distributions changes over the entire experimental session.

Importantly, the increase in the luminance level due to the stimulus onset is clearly visible in Fig. 7.4A. Yet the most important variation in the data at this level is the heart-beat and respiration related noises (see Fig. 7.4B), seen as oscillations spanning across all the recorded cortical space. The cat has approximately 4 hearth beats during the 2 seconds of stimulus presentation. These artefacts are cleaned out by computing the *fractional change.*

7.1.5 Blank Correction (Fractional Change)

Our aim is to remove the heart-beat and respiration related artefacts from the recorded neuronal data. We assume that the neuronal and artefactual data adds up linearly and that we could obtain back the neuronal data by computing the difference between the two signals which are recorded in the presence and absence of the visual stimulation. To this end, we compute the *fractional change* in order to get rid of the heart-beat related artefacts. This procedure is formally expressed in equation 7.1.

$$\frac{Condition_i(x,t) - \overline{Blank(x,t)}}{\overline{Blank(x,t)}} = \frac{Condition_i(x,t)}{\overline{Blank(x,t)}} - 1 \tag{7.1}$$

Fractional change is obtained by computing the effect of the stimulation with respect to

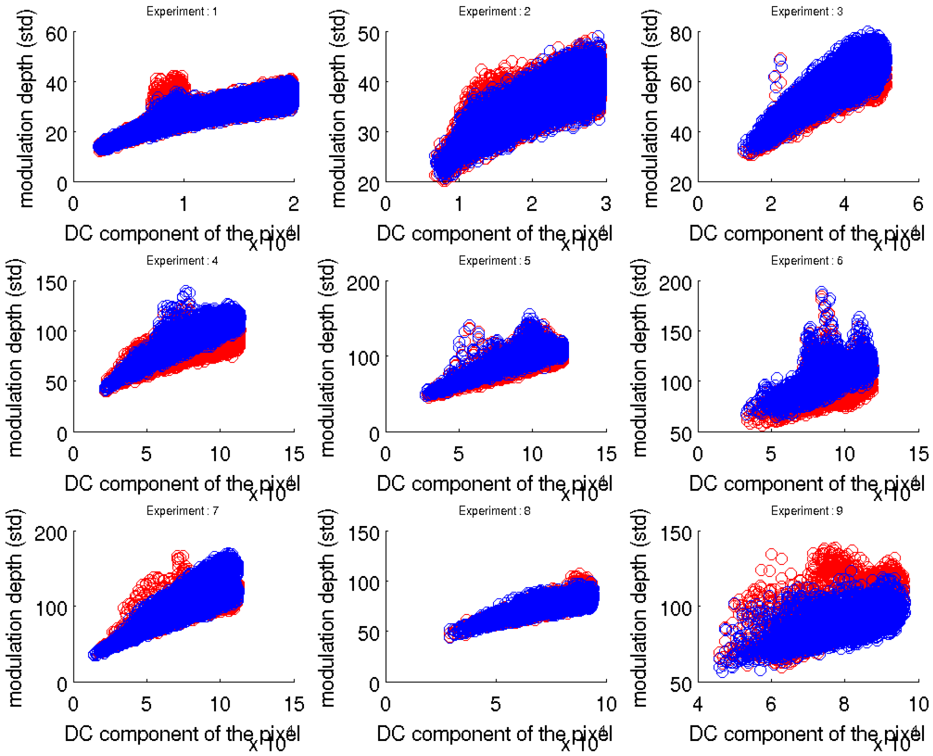

Figure 7.3: **Relationship between Prestimulus Average Luminance and Modulation Depth of a Pixel.** Each point symbolizes two quantities extracted from a pixel's luminance: DC component during prestimulus period and modulation depth, which is the standard deviation of the activity during the stimulus presentation. *Blue* and *red* dots are from grating conditions and blank conditions, respectively. Different panels depicts the data of different experiments. The presented data originates from one single trial, however same linear relationship is observed for all other trials.

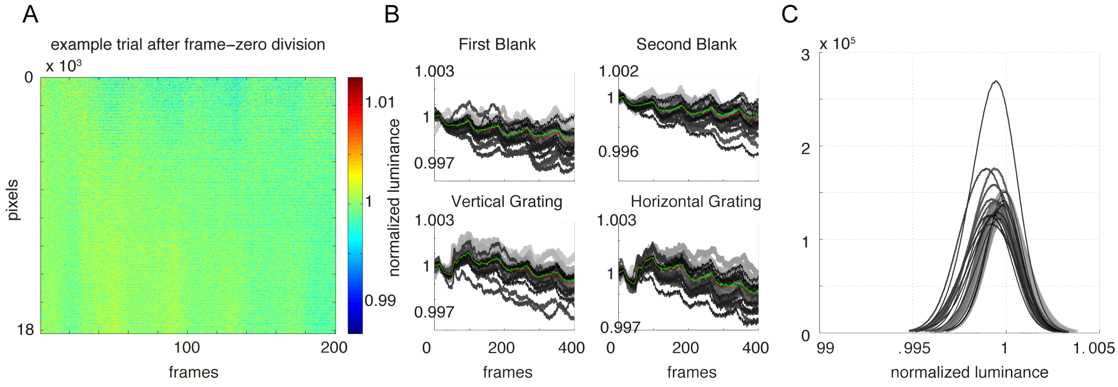

Figure 7.4: **Effect of Frame-zero Correction. (A)**Same data shown in Fig. 7.2B is presented after frame-zero correction. **(B)** The spatially averaged data is shown for all the trials recorded under 4 different conditions (panels). **(C)** Histogram of the luminance values after Frame-zero correction. Same representation as in Fig. 7.2C.

the baseline activity and a further normalization. The baseline activity is obtained from the trials where no stimulation is presented. As in each of the trialblocks, we have 2 such blank conditions, the average is taken over those before the subtraction (shown with a bar in equation 7.1). The fact that all trials are recorded in synchrony with respect to the hearth beat and respiration cycle allows us to make the subtraction method. It should be noted that the fractional change operates in an element-wise fashion. The operation is carried out in corresponding spatial and temporal recorded samples. Therefore this normalization takes into account spatial changes in luminance over the cortical surface as well as temporal non-stationariness occurring over curse of the recording.

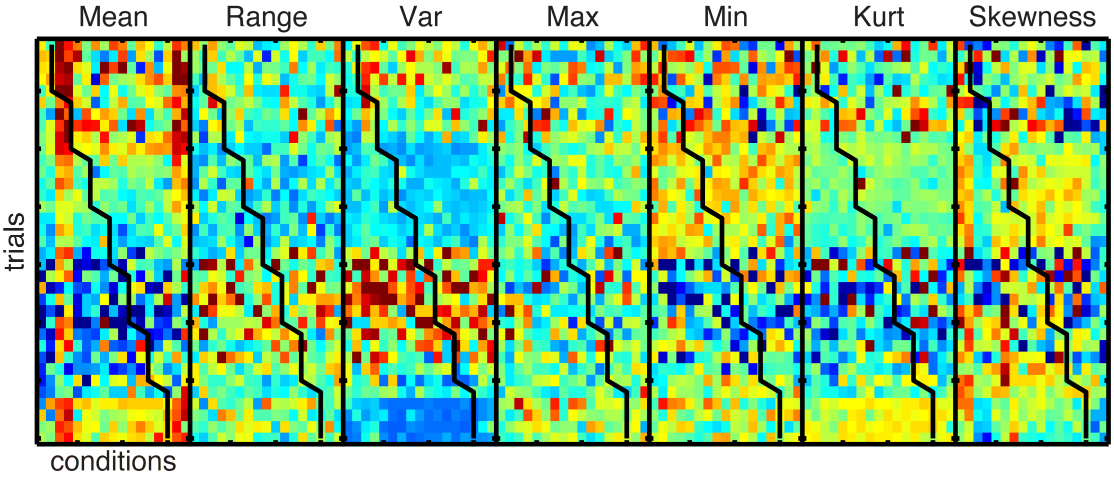

Figure 7.5: **Effect of Frame-zero and Fractional Change Operations on the Fluctuations.** The evolution of different statistics following Frame-zero correction and fractional change computations. The same representation as in Fig. 7.2. Sharp changes in the values of statistics dependent on the block identity changes is to a good extent removed.

However, it should be noted that there are a variety of choices for the definition of the baseline signal. One candidate would be the trialblock's own blank condition (this was the choice which is commonly used). In this scenario each trialblock's own blank recordings would be used for the fractional change computation. This would be the choice which is most noisy and temporarily localized. Another candidate would be the baseline computed using a given number of trials which are consecutively recorded. On the other extreme, the average blank computed across all the trials of an experiment could be also used. This would be the least noisy and temporarily located choice. These different baselines have all their pros and cons. While temporarily well-located blank recordings take into account the global fluctuations which may occur during the course of an experiment, they are also more prone to noise as the total number of trials to be based on is small. In this thesis, all the results reported, are computed using the conventional trialblock specific blanks.

In 7.5, we see the evolution of different statistics across the trials (organized exactly as in 7.2) following the operation of frame-zero correction and fractional change. In the first column which shows the mean luminance, we see a vertical stripe-like pattern at the 3rd and 4th columns. These correspond to the grating conditions which causes very strong activity levels. The first two columns shows the statistics extracted from the data recorded during blank conditions, thus in the absence of any stimulation. Important to note is that the fluctuations in the values of extracted statistics that we observed in Fig. 7.2 are to a good extent removed. Therefore these two preprocessing steps provides a good way of obtaining the specific activity levels caused by different stimulation conditions.

7.2 Behavioural Pilot Study

A behavioural pilot experiment was conducted to determine whether and how stimulus parameters affect the perceptual detectability of audiovisual congruence.

We tested different visual and auditory feature dimensions, as well as slower and faster modulation frequencies. Otherwise, audiovisual stimuli were constructed as described for the main experiment. The Gabor patch could be modulated in one of 6 dimensions: size (100-200 pixels), contrast (0.1-0.5), orientation (180 ° around horizontal axis), frequency (≈0.2-1 cycles/°), phase (0-360 °) or color saturation (0.1-0.5 along the red-green axis in DKL color space). The tone was either amplitude- (0.1-1) or frequency- (Carrier Frequency ± 40 Hz) modulated. In the latter case, the carrier frequency was chosen randomly from a uniform distribution (100-500 Hz). Modulation frequencies were bound by either 0.7, 1 or 1.3 Hz. The combinations of parameter triplets (visual, auditory, cutoff) were fully balanced.

Additionally, each combination was shown an equal number of times in congruent and incongruent stimuli. Movies were created out of 6 trajectories per modulation frequency, resulting in a total number of 432 movies. Button presses were recorded from 9 naive subjects. The whole stimulus set was divided and balanced between pairs of subjects, i.e. each subject saw 216 movies of a single balanced set. Paradigm, trial organization and instruction were identical to the main experiment (simultaneous trials). Unimodal stimuli were not shown.

On average, subjects responded correctly on 74,9 % of all trials (standard deviation: 9.3 %). The analysis of responses indicated that each subject performed well above chance (exact Binomial test, $p < 0.01$). A comparison of different visual features, auditory features and modulation frequencies revealed no significant differences (χ^2, $p < 0.05$).

We conclude that subjects are capable of distinguishing audiovisual congruent from incongruent stimuli. Furthermore, sensitivity to temporal congruence does not seem to be dependent on the specific low-level features through which the temporal structure is conveyed.

Disclaimer

All experiments reported in this thesis conforms with National and Institutional Guidelines. Experiments involving human subjects conforms with the Declaration of Helsinki.

I hereby confirm that I wrote this thesis independently and that I have not made use of resources other than those indicated. I guarantee that I significantly contributed to all materials used in this thesis. Further, this thesis was neither published in Germany nor abroad, except the parts indicated above, and has not been used to fulfill any other examination requirements. Copyright of text and figures has been or will be transfered to the respective publishers.

Osnabrück,

List of Abbreviations

A17/18	Area 17/18
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
DA	Deceleration/Acceleration Notch
EEG	Electroencephalogram
ERP	Event Related Potential
FEF	Frontal Eye Field
fMRI	Functional Magnetic Resonance Imaging
FWHM	Full Width at Half Maximum
KL	Kullback-Leibler
LC	Luminance Contrast
LFP	Local Field Potentials
LGN	Lateral Geniculus Nucleus
LIP	Lateral Intraparietal Area
MD	Modulation Depth
MEG	Magnetic Encephalogram
nSC	non-Specific Activity
PCA	Principal Component Analysis
RF	Receptive Field
RMS	Root Mean Square
ROI	Region of Interest
SC	Specific Activity
SVD	Singular Value Decomposition
V1	Primary Visual Cortex
V2	Secondary Visual Area
VSDI	Voltage-Sensitive Dye Imaging

Bibliography

Aglioti S, Cesari P, Romani M, Urgesi C (2008). Action anticipation and motor resonance in elite basketball players. *Nat Neurosci* .

Albright TD, Stoner GR (2002). Contextual influences on visual processing. *Annu Rev Neurosci* 25:339.

Allman J, Miezin F, McGuinness E (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci* 8:407.

Angelucci A, Levitt JB, Walton EJS, Hupé JM, Bullier J, Lund JS (2002). Circuits for local and global signal integration in primary visual cortex. *J Neurosci* 22(19):8633.

Anzai A, Ohzawa I, Freeman RD (1997). Neural mechanisms underlying binocular fusion and stereopsis: position vs. phase. *Proc Natl Acad Sci U S A* 94(10):5438.

Anzai A, Ohzawa I, Freeman RD (1999a). Neural mechanisms for processing binocular information i. simple cells. *J Neurophysiol* 82(2):891.

Anzai A, Ohzawa I, Freeman RD (1999b). Neural mechanisms for processing binocular information ii. complex cells. *J Neurophysiol* 82(2):909.

Arieli A, Sterkin A, Grinvald A, Aertsen A (1996). Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* 273(5283):1868.

Arndt PA, Colonius H (2003). Two stages in crossmodal saccadic integration: evidence from a visual-auditory focused attention task. *Exp Brain Res* 150(4):417.

Atick J, Li Z, Redlich A (1992). Understanding retinal color coding from first principles. *Neural computation* .

Baddeley RJ, Tatler BW (2006). High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Res* 46(18):2824.

Barlow H (1953). Summation and inhibition in the frog's retina. *J Physiol* 119(1):69.

Barlow H (1961a). *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages. The M. I. T. Press., Massachusetts Institute of Technology, (Cambridge), pp. 217–234.

Barlow HB (1961b). *Current problems in animal behaviour.*, chapter The coding of sensory messages. University Press, (Cambridge).

Basole A, White LE, Fitzpatrick D (2003). Mapping multiple features in the population response of visual cortex. *Nature* 423(6943):986.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004). Unraveling multisensory integration: patchy organization within human sts multisensory cortex. *Nat Neurosci* 7(11):1190.

Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995). Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci U S A* 92(9):3844.

Benucci A, Frazor RA, Carandini M (2007). Standing waves and traveling waves distinguish two circuits in visual cortex. *Neuron* 55(1):103.

Betsch BY, Einhäuser W, Körding KP, König P (2004). The world from a cat's perspective–statistics of natural videos. *Biol Cybern* 90(1):41.

Bidet-Caulet A, Fischer C, Bauchet F, Aguera PE, Olivier B (2007). Neural substrate of concurrent sound perception: direct electrophysiological recordings from human auditory cortex. *Front Hum Neurosci* 1(5).

Blakemore C, Campbell FW (1969). Adaptation to spatial stimuli. *J Physiol* 200(1):11P.

Blasdel GG, Salama G (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* 321(6070):579.

Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze HJ, Hillyard SA (2007). Neural basis of the ventriloquist illusion. *Curr Biol* 17(19):1697.

Bonds AB (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis Neurosci* 2(1):41.

Bonhoeffer T, Grinvald A (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* 353(6343):429.

Borg-Graham LJ, Monier C, Frégnac Y (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* 393(6683):369.

Bosking WH, Zhang Y, Schofield B, Fitzpatrick D (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J Neurosci* 17(6):2112.

Brainard DH (1997). The psychophysics toolbox. *Spat Vis* 10(4):433.

Bresciani JP, Dammeier F, Ernst MO (2008). Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events. *Brain Res Bull* 75(6):753.

Bringuier V, Chavane F, Glaeser L, Frégnac Y (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science* 283(5402):695.

Bullier J, Hupé JM, James AC, Girard P (2001). The role of feedback connections in shaping the responses of visual cortical neurons. *Prog Brain Res* 134:193.

Bushara KO, Grafman J, Hallett M (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *J Neurosci* 21(1):300.

Buzas P, Kovacs K, Ferecsko AS, Budd JML, Eysel UT, Kisvarday ZF (2006). Model-based analysis of excitatory lateral connections in the visual cortex. *J Comp Neurol* 499(6):861.

Callaway EM (1998). Local circuits in primary visual cortex of the macaque monkey. *Annu Rev Neurosci* 21:47.

Calvert GA (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb Cortex* 11(12):1110.

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10(12):2619.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997). Activation of auditory cortex during silent lipreading. *Science* 276(5312):593.

Calvert GA, Campbell R, Brammer MJ (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol* 10(11):649.

Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect. *Neuroimage* 14(2):427.

Calvert GA, Thesen T (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J Physiol Paris* 98(1-3):191.

Carandini M, Heeger DJ, Movshon JA (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci* 17(21):8621.

Caywood MS, Willmore B, Tolhurst DJ (2004). Independent components of color natural scenes resemble v1 neurons in their spatial and color tuning. *J Neurophysiol* 91(6):2859.

Celebrini S, Newsome WT (1994). Neuronal and psychophysical sensitivity to motion signals in extrastriate area mst of the macaque monkey. *J Neurosci* 14(7):4109.

Chandler DM, Field DJ (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A, Optics, image science, and vision* 24(4):922.

Chichilnisky EJ (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems* 12(2):199.

Chisum HJ, Mooser F, Fitzpatrick D (2003). Emergent properties of layer 2/3 neurons reflect the collinear arrangement of horizontal connections in tree shrew visual cortex. *J Neurosci* 23(7):2947.

Clark A (1999). An embodied cognitive science? *Trends Cogn Sci* 3(9):345.

Coppola DM, Purves HR, McCoy AN, Purves D (1998). The distribution of oriented contours in the real world. *Proc Natl Acad Sci U S A* 95(7):4002.

Corneil BD, Munoz DP (1996). The influence of auditory and visual distractors on human orienting gaze shifts. *J Neurosci* 16(24):8193.

Corneil BD, Wanrooij MV, Munoz DP, Opstal AJV (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J Neurophysiol* 88(1):438.

Creutzfeldt OD, Nothdurft HC (1978). Representation of complex visual stimuli in the brain. *Naturwissenschaften* 65(6):307.

Cumming BG (2002). An unexpected specialization for horizontal disparity in primate primary visual cortex. *Nature* 418(6898):633.

Cumming BG, DeAngelis GC (2001). The physiology of stereopsis. *Annu Rev Neurosci* 24:203.

Dan Y, Atick JJ, Reid RC (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci* 16(10):3351.

David SV, Gallant JL (2005). Predicting neuronal responses during natural vision. *Network* 16(2-3):239.

David SV, Vinje WE, Gallant JL (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci* 24(31):6991.

de Ruyter van Steveninck RR, Lewen GD, Strong SP, Koberle R, Bialek W (1997). Reproducibility and variability in neural spike trains. *Science* 275(5307):1805.

Dean AF, Tolhurst DJ (1983). On the distinctness of simple and complex cells in the visual cortex of the cat. *J Physiol* 344:305.

DeAngelis G (2000). Seeing in three dimensions: the neurophysiology of stereopsis. *Trends Cogn Sci* 4(3):80.

Dehay C, Kennedy H, Bullier J (1988). Characterization of transient cortical projections from auditory, somatosensory, and motor cortices to visual areas 17, 18, and 19 in the kitten. *J Comp Neurol* 272(1):68.

Delorme A, Makeig S (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J Neurosci Methods*

134(1):9.

Dhamala M, Assisi CG, Jirsa VK, Steinberg FL, Kelso JAS (2007). Multisensory integration for timing engages different brain networks. *Neuroimage* 34(2):764.

Ding J, Sperling G, Srinivasan R (2006). Attentional modulation of ssvep power depends on the network tagged by the flicker frequency. *Cereb Cortex* 16(7):1016.

Doehrmann O, Naumer M (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Res* .

Douglas RJ, Martin KAC (2004). Neuronal circuits of the neocortex. *Annu Rev Neurosci* 27:419.

Douglas RJ, Martin KAC (2007). Mapping the matrix: the ways of neocortex. *Neuron* 56(2):226.

Driver J, Noesselt T (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron* 57(1):11.

Driver J, Spence C (2000). Multisensory perception: beyond modularity and convergence. *Curr Biol* 10(20):R731.

Einhäuser W, Hipp J, Eggert J, Korner E, König P (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biol Cybern* 93(1):79.

Einhäuser W, Kayser C, König P, Körding KP (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur J Neurosci* 15(3):475.

Einhäuser W, Kayser C, Körding KP, König P (2003). Learning distinct and complementary feature selectivities from natural colour videos. *Rev Neurosci* 14(1-2):43.

Einhäuser W, König P (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *Eur J Neurosci* 17(5):1089.

Einhäuser W, Moeller GU, Condradt J, Vockeroth J, Bartl K, Schneider E, König P (2008). Eye-head coordination during free exploration in human and cat. *Ann New York Acad Sci* .

Escabi MA, Miller LM, Read HL, Schreiner CE (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J Neurosci* 23(37):11489.

Ethier C, Brizzi L, Darling WG, Capaday C (2006). Linear summation of cat motor cortex outputs. *J Neurosci* 26(20):5574.

Felleman DJ, Essen DCV (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1(1):1.

Felsen G, Dan Y (2005). A natural approach to studying vision. *Nat Neurosci* 8(12):1643.

Felsen G, Touryan J, Dan Y (2005a). Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. *Network* 16(2-3):139.

Felsen G, Touryan J, Han F, Dan Y (2005b). Cortical sensitivity to visual features in natural scenes. *PLoS Biol* 3(10):e342.

Ferezou I, Hill EL, Cauli B, Gibelin N, Kaneko T, Rossier J, Lambolez B (2007). Extensive overlap of mu-opioid and nicotinic sensitivity in cortical interneurons. *Cereb Cortex* 17(8):1948.

Field DJ (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A* 4(12):2379.

Fitzpatrick D (2000). Seeing beyond the receptive field in primary visual cortex. *Curr Opin Neurobiol* 10(4):438.

Földíak P (1991). Learning invariance from transformation sequences. *Neural Comput* 3(2):194.

Foxe JJ, Schroeder CE (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport* 16(5):419.

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

Freeman WJ (2000). *How brains make up their minds.* Columbia University Press, New York.

Frégnac Y, Shulz DE (1999). Activity-dependent regulation of receptive field properties of cat area 17 by supervised hebbian learning. *J Neurobiol* 41(1):69.

Fu KMG, Johnston TA, Shah AS, Arnold L, Smiley J, Hackett TA, Garraghty PE, Schroeder CE (2003). Auditory cortical neurons respond to somatosensory stimulation. *J Neurosci* 23(20):7510.

Galambos R, Makeig S, Talmachoff PJ (1981). A 40-hz auditory potential recorded from the human scalp. *Proc Natl Acad Sci U S A* 78(4):2643.

Geisler WS, Albrecht DG, Crane AM (2007). Responses of neurons in primary visual cortex to transient changes in local contrast and luminance. *J Neurosci* 27(19):5063.

Georgopoulos AP (1994). New concepts in generation of movement. *Neuron* 13(2):257.

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J Neurosci* 25(20):5004.

Giard MH, Peronnet F (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J Cogn Neurosci* 11(5):473.

Gibson JJ, Radner M (1937). Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of Experimental Psychology* 12:453. Notes.

Gilbert CD, Wiesel TN (1989). Columnar specificity of intrinsic horizontal and cortico-cortical connections in cat visual cortex. *J Neurosci* 9(7):2432.

Glimcher PW (2003). *Decisions, uncertainty, and the brain the science of neuroeconomics.* MIT Press, Cambridge, Mass.

Graham DJ, Field DJ, Squire LR (2009). *Natural Images: Coding Efficiency.* Academic Press, Oxford, pp. 19–27.

Grinvald A, Anglister L, Freeman JA, Hildesheim R, Manker A (1984). Real-time optical imaging of naturally evoked electrical activity in intact frog brain. *Nature* 308(5962):848.

Grinvald A, Hildesheim R (2004). Vsdi: a new era in functional imaging of cortical dynamics. *Nat Rev Neurosci* 5(11):874.

Grinvald A, Lieke EE, Frostig RD, Hildesheim R (1994). Cortical point-spread function and long-range lateral interactions revealed by real-time optical imaging of macaque monkey primary visual cortex. *J Neurosci* 14(5 Pt 1):2545.

Gross CG, Rocha-Miranda CE, Bender DB (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J Neurophysiol* 35(1):96.

Held R, Hein A (1963). Movement-produced stimulation in the development of visually guided behavior. *J Comp Physiol Psychol* 56:872.

Herculano-Houzel S (2009). The human brain in numbers: A linearly scaled-up primate brain. *Front Hum Neurosci* 3:31.

Hirsch JA, Alonso JM, Reid RC, Martinez LM (1998). Synaptic integration in striate cortical simple cells. *J Neurosci* 18(22):9517.

Hoyer PO, Hyvarinen A (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network* 11(3):191.

Hubel D, Wiesel T (1962a). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106.

Hubel D, Wiesel T (1962b). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106.

Hubel D, Wiesel TN (1959). Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574.

Hubel DH, Wiesel TN (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol* 28:229.

Hubel DH, Wiesel TN (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *J Comp Neurol* 158(3):267.

Hubener M, Shoham D, Grinvald A, Bonhoeffer T (1997). Spatial relationships among three columnar systems in cat area 17. *J Neurosci* 17(23):9270.

Humphrey AL, Sur M, Uhlrich DJ, Sherman SM (1985). Termination patterns of individual x- and y-cell axons in the visual cortex of the cat: projections to area 18, to the 17/18 border region, and to both areas 17 and 18. *J Comp Neurol* 233(2):190.

Hupé JM, James AC, Girard P, Bullier J (2001). Response modulations by static texture surround in area v1 of the macaque monkey do not depend on feedback connections from v2. *J Neurophysiol* 85(1):146.

Hurri J, Hyvärinen A (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural computation* 15(3):663.

Hyvärinen A, Hoyer PO (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision research* 41(18):2413.

Ito M, Tamura H, Fujita I, Tanaka K (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology* 73(1):218.

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

Jackendoff R (1987). *Consciousness and the computational mind.* MIT Press, Cambridge, Mass.

Jancke D (2000). Orientation formed by a spot's trajectory: a two-dimensional population approach in primary visual cortex. *J Neurosci* 20(14):RC86.

Jancke D, Erlhagen W, Schoner G, Dinse HR (2004). Shorter latencies for motion trajectories than for flashes in population responses of cat primary visual cortex. *The Journal of Physiology Online* 556(3):971. 10.1113/jphysiol.2003.058941.

Jones HE, Grieve KL, Wang W, Sillito AM (2001). Surround suppression in primate v1. *J Neurophysiol* 86(4):2011.

Julesz B (1960). Binocular depth perception of computer-generated patterns. *The Bell Systems Technical Journal* .

Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski TJ (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clin Neurophysiol* 111(10):1745.

Kanwisher N, McDermott J, Chun MM (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* 17(11):4302.

Kapadia MK, Ito M, Gilbert CD, Westheimer G (1995). Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in v1 of alert monkeys. *Neuron* 15(4):843.

Kapadia MK, Westheimer G, Gilbert CD (1999). Dynamics of spatial summation in primary visual cortex of alert monkeys. *Proc Natl Acad Sci U S A* 96(21):12073.

Kayser C, König P (2004). Stimulus locking and feature selectivity prevail in complementary frequency ranges of v1 local field potentials. *Eur J Neurosci* 19(2):485.

Kayser C, Körding KP, König P (2003a). Learning the nonlinearity of neurons from natural visual stimuli. *Neural Comput* 15(8):1751.

Kayser C, Körding KP, König P (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Curr Opin Neurobiol* 14(4):468.

Kayser C, Logothetis NK (2007). Do early sensory cortices integrate cross-modal information? *Brain Struct Funct* 212(2):121.

Kayser C, Petkov CI, Augath M, Logothetis NK (2005). Integration of touch and sound in auditory cortex. *Neuron* 48(2):373.

Kayser C, Petkov CI, Augath M, Logothetis NK (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *J Neurosci* 27(8):1824. FMRI, monkey stimuli: wildlife movies with sounds.

Kayser C, Petkov CI, Logothetis NK (2008). Visual modulation of neurons in auditory cortex. *Cereb Cortex* 18(7):1560.

Kayser C, Salazar RF, König P (2003b). Responses to natural scenes in cat v1. *J Neurophysiol* 90(3):1910.

Kisvarday ZF, Kim DS, Eysel UT, Bonhoeffer T (1994). Relationship between lateral inhibitory connections and the topography of the orientation map in cat visual cortex. *Eur J Neurosci* 6(10):1619.

Koch C, Ullman S (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol* 4(4):219.

Kohn A, Movshon JA (2003). Neuronal adaptation to visual motion in area mt of the macaque. *Neuron* 39(4):681.

Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007). Causal inference in multisensory perception. *PLoS ONE* 2(9):e943.

Körding KP, Kayser C, Einhäuser W, König P (2004). How are complex cell properties adapted to the statistics of natural stimuli? *J Neurophysiol* 91(1):206.

Körding KP, König P (2001). Supervised and unsupervised learning with two sites of synaptic integration. *J Comput Neurosci* 11(3):207.

Krekelberg B, Boynton GM, van Wezel RJA (2006). Adaptation: from single cells to bold signals. *TRENDS in Neurosciences* 29(5):250.

Kuffler SW (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16(1):37.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53(2):279.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320(5872):110.

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94(3):1904.

Lamme VA, Roelfsema PR (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* 23(11):571.

Larkum ME, Zhu JJ, Sakmann B (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* 398(6725):338.

Laughlin S (1981). A simple coding procedure enhances a neuron's information capacity.

Z Naturforsch C 36(9-10):910.

Laughlin SB, de Ruyter van Steveninck RR, Anderson JC (1998). The metabolic cost of neural information. *Nat Neurosci* 1(1):36.

Lee SH, Blake R (1999). Visual form created solely from temporal structure. *Science* 284(5417):1165.

Lennie P (2003). The cost of cortical computation. *Curr Biol* 13(6):493.

Lesica NA, Jin J, Weng C, Yeh CI, Butts DA, Stanley GB, Alonso JM (2007). Adaptation to stimulus contrast and correlations during natural visual stimulation. *Neuron* 55(3):479.

LeVay S, Connolly M, Houde J, Van Essen DC (1985). The complete pattern of ocular dominance stripes in the striate cortex and visual field of the macaque monkey. *J Neurosci* 5(2):486.

Levitt JB, Lund JS (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature* 387(6628):73.

Liu J, Newsome WT (2006). Local field potential in cortical area mt: stimulus tuning and behavioral correlations. *J Neurosci* 26(30):7779.

Macaluso E, Driver J (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci* 28(5):264.

Macaluso E, Frith CD, Driver J (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science* 289(5482):1206.

Macaluso E, George N, Dolan R, Spence C, Driver J (2004). Spatial and temporal factors during processing of audiovisual speech: a pet study. *Neuroimage* 21(2):725.

Maffei L, Fiorentini A (1976). The unresponsive regions of visual cortical receptive fields. *Vision Res* 16(10):1131.

Maffei L, Fiorentini A, Bisti S (1973). Neural correlate of perceptual adaptation to gratings. *Science* 182(116):1036.

Mainen ZF, Sejnowski TJ (1995). Reliability of spike timing in neocortical neurons. *Science* 268(5216):1503.

Malach R, Amir Y, Harel M, Grinvald A (1993). Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proc Natl Acad Sci U S A* 90(22):10469.

Mante V, Carandini M (2005). Mapping of stimulus energy in primary visual cortex. *J Neurophysiol* 94(1):788.

Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nat Neurosci* 8(12):1690.

Marr D (1982). *Vision : a computational investigation into the human representation and processing of visual information.* W.H. Freeman, San Francisco.

Martinez LM, Wang Q, Reid RC, Pillai C, Alonso JM, Sommer FT, Hirsch JA (2005). Receptive field structure varies with layer in the primary visual cortex. *Nat Neurosci* 8(3):372. 10.1038/nn1404.

Marzetti L, Gratta CD, Nolte G (2008). Understanding brain connectivity from eeg data by identifying systems composed of interacting sources. *Neuroimage* .

Mazzoni A, Panzeri S, Logothetis NK, Brunel N (2008). Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons. *PLoS Comput Biol* 4(12):e1000239.

McBeath MK, Shaffer DM, Kaiser MK (1995). How baseball outfielders determine where to run to catch fly balls. *Science* 268(5210):569.

McLaughlin D, Shapley R, Shelley M, Wielaard DJ (2000). A neuronal network model of macaque primary visual cortex (v1): orientation selectivity and dynamics in the input layer 4calpha. *Proc Natl Acad Sci U S A* 97(14):8087.

McLeod P, Dlenes Z (1993). Running to catch the ball. *Nature* 362(6415):23.

Meredith MA, Nemitz JW, Stein BE (1987). Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *J Neurosci* 7(10):3215.

Meredith MA, Stein BE (1983). Interactions among converging sensory inputs in the superior colliculus. *Science* 221(4608):389.

Meredith MA, Stein BE (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol* 56(3):640.

Miller LM, D'Esposito M (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J Neurosci* 25(25):5884.

Mishra J, Martinez A, Hillyard SA (2008). Cortical processes underlying sound-induced flash fusion. *Brain Res* 1242:102.

Mishra J, Martinez A, Sejnowski TJ, Hillyard SA (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *J Neurosci* 27(15):4120.

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res Cogn Brain Res* 14(1):115.

Monier C, Chavane F, Baudot P, Graham LJ, Frégnac Y (2003). Orientation and direction selectivity of synaptic inputs in visual cortical neurons: A diversity of combinations produces spike tuning. *Neuron* 37(4):663.

Movshon JA, Lennie P (1979). Pattern-selective adaptation in visual cortical neurones. *Nature* 278(5707):850.

Muller MM, Malinowski P, Gruber T, Hillyard SA (2003). Sustained division of the attentional spotlight. *Nature* 424(6946):309.

Nagel S, Carl C, Kringe T, Martin R, König P (2005). Beyond sensory substitution—learning the sixth sense. *Journal of Neural Engineering* .

Nelson JI, Frost BJ (1985). Intracortical facilitation among co-oriented, co-axially aligned simple cells in cat striate cortex. *Exp Brain Res* 61(1):54.

Noë A (2004). *Action in perception.* MIT Press, Cambridge, Mass.

Noesselt T, Bonath B, Boehler CN, Schoenfeld MA, Heinze HJ (2008). On perceived synchrony-neural dynamics of audiovisual illusions and suppressions. *Brain Res* 1220:132.

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze HJ, Driver J (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J Neurosci* 27(42):11431.

Nunez PL (1981). *Electric fields of the brain : the neurophysics of EEG.* Oxford University Press, New York.

Ohzawa I, DeAngelis GC, Freeman RD (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* 249(4972):1037.

Ohzawa I, Sclar G, Freeman RD (1982). Contrast gain control in the cat visual cortex. *Nature* 298(5871):266.

O'Keefe J, Dostrovsky J (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Res* 34(1):171.

Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607.

Onat S, Jancke D, König P (2009a). Processing of locally presented natural movies and contextual interactions under natural conditions. *in Preparation* .

Onat S, König P, Jancke D (2009b). Multiplexing information about position and orientation in early visual cortex: A voltage-sensitive dye imaging study. *Submitted* .

Onat S, König P, Jancke D (2009c). Processing of natural movies revealed by voltage-sensitive dye imaging across primary visual cortex. *Submitted* .

Onat S, Libertus K, König P (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision* 7(10):11.1.

O'Regan JK, Noë A (2001). A sensorimotor account of vision and visual consciousness. *Behav Brain Sci* 24(5):939.

Palva S, Palva JM (2007). New vistas for alpha-frequency band oscillations. *Trends Neurosci* 30(4):150.

Parkhurst D, Law K, Niebur E (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Res* 42(1):107.

Pelli DG (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis* 10(4):437.

Pestilli F, Viera G, Carrasco M (2007). How do attention and adaptation affect contrast sensitivity? *J Vis* 7(7):9.1.

Peters RJ, Iyer A, Itti L, Koch C (2005). Components of bottom-up gaze allocation in natural images. *Vision Res* 45(18):2397.

Petersen CCH, Grinvald A, Sakmann B (2003a). Spatiotemporal dynamics of sensory responses in layer 2/3 of rat barrel cortex measured in vivo by voltage-sensitive dye imaging combined with whole-cell voltage recordings and neuron reconstructions. *J Neurosci* 23(4):1298.

Petersen CCH, Grinvald A, Sakmann B (2003b). Spatiotemporal dynamics of sensory responses in layer 2/3 of rat barrel cortex measured in vivo by voltage-sensitive dye imaging combined with whole-cell voltage recordings and neuron reconstructions. *J Neurosci* 23(4):1298.

Pettigrew JD, Nikara T, Bishop PO (1968). Binocular interaction on single units in cat striate cortex: simultaneous stimulation by single moving slit with receptive fields in correspondence. *Exp Brain Res* 6(4):391.

Picton TW, Skinner CR, Champagne SC, Kellett AJ, Maiste AC (1987). Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. *J Acoust Soc Am* 82(1):165.

Poggio GF, Fischer B (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J Neurophysiol* 40(6):1392.

Polat U, Mizobe K, Pettet MW, Kasamatsu T, Norcia AM (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature* 391(6667):580.

Prechtl JC, Cohen LB, Pesaran B, Mitra PP, Kleinfeld D (1997). Visual stimuli induce waves of electrical activity in turtle cortex. *Proc Natl Acad Sci U S A* 94(14):7621.

Priebe NJ, Ferster D (2008). Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron* 57(4):482.

Rager G, Singer W (1998). The response of cat visual cortex to flicker stimuli of variable frequency. *Eur J Neurosci* 10(5):1856.

Read JCA, Parker AJ, Cumming BG (2002). A simple model accounts for the response of disparity-tuned v1 neurons to anticorrelated images. *Vis Neurosci* 19(6):735.

Regan D (1966). Some characteristics of average steady-stae and transient responses evoked by modulated light. *Electroceph clin Neurophysiology* 20:238.

Reinagel P, Zador AM (1999). Natural scene statistics at the centre of gaze. *Network* 10(4):341.

Renninger LW, Coughlan J, Verghese P, Malik J (2005). An information maximization model of eye movements. *Adv Neural Inf Process Syst* 17:1121.

Ringach DL (2004). Mapping receptive fields in primary visual cortex. *J Physiol* 558(Pt 3):717.

Rockland KS, Lund JS (1982). Widespread periodic intrinsic connections in the tree shrew visual cortex. *Science* 215(4539):1532.

Roland PE, Hanazawa A, Undeman C, Eriksson D, Tompa T, Nakamura H, Valentiniene S, Ahmed B (2006). Cortical feedback depolarization waves: a mechanism of top-down influence on early visual areas. *Proc Natl Acad Sci U S A* 103(33):12586.

Rolls ET, Deco G (2006). Attention in natural scenes: Neurophysiological and computational bases. *Neural Netw* 19(9):1383.

Ruderman D, Bialek W (1994). Statistics of natural images: Scaling in the woods. *Phys Rev Lett* 73(6):814.

Rust NC, Movshon JA (2005). In praise of artifice. *Nat Neurosci* 8(12):1647.

Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron* 46(6):945.

Schall S (2008). Amplitude locking of eeg activity to dynamic audiovisual stimuli. *Master Thesis* .

Schall S, Quigley C, Onat S, König P (2009). Visual stimulus locking of eeg is modulated by temporal congruency of auditory stimuli. *Experimental brain research Experimentelle Hirnforschung Experimentation cerebrale* .

Schnupp JW, Mrsic-Flogel TD, King AJ (2001). Linear processing of spatial cues in primary auditory cortex. *Nature* 414(6860):200.

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* 12(3):106.

Sengpiel F, Sen A, Blakemore C (1997). Characteristics of surround inhibition in cat area 17. *Exp Brain Res* 116(2):216.

Senkowski D, Schneider TR, Foxe JJ, Engel AK (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci* 31(8):401.

Senkowski D, Talsma D, Grigutsch M, Herrmann CS, Woldorff MG (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia* 45(3):561.

Seriès P, Lorenceau J, Frégnac Y (2003). The ”silent” surround of v1 receptive fields: theory and experiments. *J Physiol Paris* 97(4-6):453.

Shams L, Kamitani Y, Shimojo S (2000). Illusions. what you see is what you hear. *Nature* 408(6814):788.

Shams L, Kamitani Y, Shimojo S (2002). Visual illusion induced by sound. *Brain research Cognitive brain research* 14(1):147.

Shapley R, Hawken M, Ringach DL (2003). Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron* 38(5):689.

Sharon D, Grinvald A (2002). Dynamics and constancy in cortical spatiotemporal patterns of orientation processing. *Science* 295(5554):512.

Sharon D, Jancke D, Chavane F, Na'aman S, Grinvald A (2007). Cortical response field dynamics in cat visual cortex. *Cereb Cortex* 17(12):2866.

Shipley T (1964). Auditory flutter-driving of visual flicker. *Science* 145:1328.

Shoham D, Glaser DE, Arieli A, Kenet T, Wijnbergen C, Toledo Y, Hildesheim R, Grinvald A (1999). Imaging cortical dynamics at high spatial and temporal resolution with novel blue voltage-sensitive dyes. *Neuron* 24(4):791.

Sillito AM, Grieve KL, Jones HE, Cudeiro J, Davis J (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* 378(6556):492.

Simoncelli EP, Olshausen BA (2001). Natural image statistics and neural representation. *Annu Rev Neurosci* 24:1193.

Smyth D, Willmore B, Baker GE, Thompson ID, Tolhurst DJ (2003). The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *J Neurosci* 23(11):4746.

Somers DC, Nelson SB, Sur M (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *J Neurosci* 15(8):5448.

Sprekeler H, Michaelis C, Wiskott L (2007). Slowness: an objective for spike-timing-dependent plasticity? *PLoS Computational Biology* 3(6):e112.

Stanford TR, Quessy S, Stein BE (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J Neurosci* 25(28):6499.

Stein BE, Meredith MA (1993). *The merging of the senses.* MIT Press, Cambridge, Mass.

Steriade M, Gloor, P, Llinas RR, da Silva FHL, Mesulam MM (1990). Basic mechanisms of cerbral rhythmic activities. *Electroceph clin Neurophysiology* 76(6):481.

Sterkin A, Lampl I, Ferster D, Grinvald A, Arieli A (1998). Real time optical imaging in cat visual cortex exhibits high similarity to intracellular activity. *Neurosci Lett* 51(S41).

Tallon-Baudry C, Bertrand O (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn Sci* 3(4):151.

Tatler BW, Baddeley RJ, Gilchrist ID (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Res* 45(5):643.

Tatler BW, Baddeley RJ, Vincent BT (2006). The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vision Res* 46(12):1857.

Thomas OM, Cumming BG, Parker AJ (2002). A specialization for relative disparity in v2. *Nat Neurosci* 5(5):472.

Torralba A (2003). Modeling global scene factors in attention. *J Opt Soc Am A Opt Image Sci Vis* 20(7):1407.

Toth LJ, Rao SC, Kim DS, Somers D, Sur M (1996). Subthreshold facilitation and suppression in primary visual cortex revealed by intrinsic signal imaging. *Proc Natl Acad Sci U S A* 93(18):9869.

Touryan J, Lau B, Dan Y (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *J Neurosci* 22(24):10811.

Tucker TR, Katz LC (2003). Spatiotemporal patterns of excitation and inhibition evoked by the horizontal network in layer 2/3 of ferret visual cortex. *J Neurophysiol* 89(1):488.

van Atteveldt NM, Formisano E, Blomert L, Goebel R (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb Cortex* 17(4):962.

van der Schaaf A, van Hateren JH (1996). Modelling the power spectra of natural images:

statistics and information. *Vision Res* 36(17):2759.

Varela FJ, Thompson ET, Rosch E (1992). *The embodied mind cognitive science and human experience.* MIT Press, Cambridge, Mass.

Vautin RG, Berkley MA (1977). Responses of single cells in cat visual cortex to prolonged stimulus movement: neural correlates of visual aftereffects. *J Neurophysiol* 40(5):1051.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al (2001). The sequence of the human genome. *Science* 291(5507):1304.

Vinje WE, Gallant JL (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287(5456):1273.

Vinje WE, Gallant JL (2002). Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *J Neurosci* 22(7):2904.

von Stein A, Chiang C, König P (2000). Top-down processing mediated by interareal synchronization. *Proc Natl Acad Sci U S A* 97(26):14748.

Vroomen J, de Gelder B (2004). *Perceptual effects of cross-modal stimulation: The cases of ventriloquism and the freezing phenomenon.*

Walker GA, Ohzawa I, Freeman RD (1999). Asymmetric suppression outside the classical receptive field of the visual cortex. *J Neurosci* 19(23):10536.

Wallace MT, Wilkinson LK, Stein BE (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *J Neurophysiol* 76(2):1246.

Wickens TD (2002). *Elementary signal detection theory.* Oxford University Press, Oxford; New York.

Wiesel TN, Hubel DH (1974). Ordered arrangement of orientation columns in monkeys lacking visual experience. *J Comp Neurol* 158(3):307.

Wiskott L, Sejnowski TJ (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput* 14(4):715.

Wyss R, König P, Verschure PFMJ (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol* 4(5):e120.

Xu W, Huang X, Takagaki K, young Wu J (2007). Compression and reflection of visually evoked cortical waves. *Neuron* 55(1):119.